Australian Army
Research Centre

Accelerated Preparedness—Scalability Insights for Defence

Renée Kidson

Australian Army Occasional Paper No. 22

Australian Army
Research Centre

Accelerated Preparedness— Scalability Insights for Defence

Renée Kidson

Australian Army Occasional Paper No. 22

Serving the Nation

The views expressed in this publication are those of the author(s) and do not necessarily reflect the official policy or position of the Australian Army, the Department of Defence or the Australian Government.

ISSN (Print) 2653-0406
ISSN (Digital) 2653-0414

DOI: https://doi.org/10.61451/267505

All enquiries regarding this publication should be forwarded to the Director of the Australian Army Research Centre.

To learn about the work of the Australian Army Research Centre visit researchcentre.army.gov.au

Cover image: Indian and Australian Army soldiers share infantry tactics, techniques and procedures during Exercise Austrahind 22. (Source: Defence image gallery)

Executive Summary

Prepare to Scale

Scalability is about how an organisation's performance responds to significant changes in workload. The workload may be changing in *quantity* (more, or less, of the same) or *type* (existing products and services, or new ones), challenging the current size and shape of an organisation.

Sound familiar? Recent events (e.g. the COVID-19 pandemic) have forced many organisations—public, private and for-purpose—to rapidly scale: upwards, downwards, inwards, outwards. And this has often been without notice or warning. For the Australian Defence Force (ADF), the 2023 Defence Strategic Review[1] underscores the vanishing notion of strategic warning time and calls for Defence to undertake 'accelerated preparedness'. This is a strong cue to examine scalability within the Defence organisation.

The term 'scalability' is often thrown around in executive parlance. But in commencing this journey, I quickly appreciated that scalability—as a concept, as theory and as practitioner guidance—did not exist in coherent form for Defence, or for any other organisation for that matter. This paper steps into that breach.

The key premise of this work is that an organisation can enhance its scalability. It can do this through sound scalability design, and through impactful scalability response. But scalability is a *craft*—combining both science (requiring technical expertise) and the art form of 'knowing your business'. It is also premised on understanding changes in the operating environment. Perfecting a craft requires some knowledge, and some practice—in advance of the need to perform it. This paper provides the foundation knowledge necessary to start the scalability journey, whether your business is warfighting, leading a public agency, turning a profit or running a charity. Given the challenges facing Defence to scale in response to rapidly changing strategic circumstances, the insights in this paper can provide assurance of organisational resilience to those charged with leading change.

The scalability leadership task is to ‘find and fix’ the sequence of binding constraints that any scaling response will encounter. This applies whether you are leading at an organisational level or leading individual business processes. Regardless, as a leader you will need a scalability mindset, sense-makers and a scaling strategy to do this. You will need to know the difference between first and second-order scalability, and you will certainly need to know how your organisation creates value, and about its capacity components. This paper shares these and other facets of scalability.

Specifically:

- **Part 1** presents scalability theory—defining what scalability is, describing how to scale (methods), and developing an initial conceptual model of scalability.
- **Part 2** is practitioner focused, presenting:
    - the scoping and planning considerations for real people, teams and organisations directed to scale
    - an industry case study of first-order scalability, based on the Australian retail supermarket response to the COVID-19 pandemic
    - scaling principles and metrics for benchmarking and reporting
    - scalability implications for Australian military strategy and capabilities.

Originally written to identify the scalability implications for the ADF, this paper starts from theoretical scratch, achieves that original practitioner's objective, and then goes considerably beyond, presenting the definitive current 'state of the art' for scalability. For practitioners, the work contains handy flick-through tools, including:

- 'Scalability Action Plan-on-a-Page for Leaders' as an aide mémoire
- 'Quick Take-Outs' for each chapter
- 'So What for Defence? The Scalability Top 10'
- illustrations to explain scalability concepts.

For the ADF specifically, this work challenges the organisation to identify those capabilities that deliver both asymmetric results and scalability in the context of military strategy. Interested? Read on. Then over to you, to implement scalability in your own business process, system or organisation!

Scalability Action Plan-on-a-Page for Leaders

If you are asked to scale, you need to determine three parameters:

a. **Scaling factor**: e.g. x 2, x 4

   i). In multiples of original state

   ii). *What size?*

b. **Scaling rate**: time required to double your throughput, outputs or effects

   i). *How fast?* [to achieve the scaling effect]

c. **Scaling ratio**: shape of the scaled outcome

   i). Core to enabling (functions, services, capabilities etc.)

   ii). *What shape?*

There are three steps to a scaling response:

1. **Scope and frame**:

   a. ***Why?*** Clarify scaling intention, and scaling imperative (internal driver or external driver?)

   b. ***What?*** Are you scaling:

      i). an existing or new capability?

      ii). hardware (e.g. assets and infrastructure) or software (people and processes)?

   c. ***When?***

      i). Urgency: *How fast?* (scaling rate)

      ii). Duration: *How long?* (expediency or sustainability)

   d. ***Where?*** Centralised or decentralised? Control measures?

   e. ***Who?*** Who are your:

      i). sponsors?

      ii). targets?

      iii). enabling personnel?

      iv). commentators?

    f. ***How?***

        i). Horizontal or vertical scaling method?

        ii). Do you need a state transition (from first- to second-order scalability) to unlock more performance?

2. **Plan**:

    a. Identify the expected sequence of binding constraints.

    b. Define what is essential and what is non-essential in your organisation (contingent on the specific scaling scenario).

    c. Develop scaling options—managing both supply and demand (for specific products and services).

    d. Evaluate scaling options (according to criteria which may include performance, cost, reliability, security, timeliness).

    e. Select and implement the preferred option(s).

    f. Measure and monitor scaling effects.

3. **Execute**:

    a. Redirect existing capacity from non-essential to essential, dependent on the scaling imperative.

    b. Harness latent capacity. Your business-as-usual inefficiency is necessary redundancy for scaling.

    c. Navigate the scalability–complexity trade-off. Simpler capabilities are easier to scale. Are they an effective first-order response?

    d. Stay in shape! There is risk of disproportional scaling (e.g. of core to enabling capabilities). Maintain positive control by:

        i). ruthless pursuit of the binding constraint

        ii). identification and amplification of critical enablers

        iii). calibrating the scaling response to the absorptive capacity of the external operating environment.

    e. Exploit scaling as a transformation opportunity. Accelerate and embed positive change.

Contents

Introduction

Recent rapid developments in the global, regional and domestic environments have increasingly challenged governments, businesses and organisations at all levels to respond with scalable solutions. For the Australian Defence Force (ADF), the last several years have seen unprecedented demands for domestic humanitarian aid and disaster relief contingencies, ranging from the national bushfire emergency (2019–2020), to the COVID-19 pandemic, record-breaking floods in northern New South Wales and southern Queensland, and the aged care crisis. Although these recent contingencies fall largely within the category of 'non-traditional' security threats, the ADF has nonetheless been required to respond at scale.

The term 'scalability' is becoming more frequently used, as a (presumably desirable) attribute of teams, processes and organisations. However, conceptual development of scalability in the literature has been rudimentary thus far, emerging in isolation in technical and social science disciplines but lacking an integrated, trans-disciplinary perspective. Consequently, the term lacks precision in application, with vague and varied definitions of 'scalability' and how to achieve it in specific contexts. In a military context, unlike the related concept of mobilisation, scalability is currently undefined and is not yet included in doctrine.

This paper seeks to make a new and innovative contribution to the broad field of organisational theory and design by developing the theory and practice components of scalability as a concept. For the purpose of this work, 'organisation' is considered to include both public and business enterprises. The work is presented in two parts. Part 1 develops scalability theory; and Part 2 presents industry practitioner perspectives on scalability, applying both the theory and these industry insights to Defence and, within it, the ADF.

This paper is based on an important premise: that an organisation can enhance its scalability.

Intended Audience

The intended audience for this work, in the first instance, is Defence decision-makers, force designers and planners, practitioners, and those with an interest in strategic mobilisation. This paper is intended as a catalyst for debate, and as an initial aid to thinking about scalability in an organisational context. This work aims to equip practitioners with concepts, principles and a lexicon to undertake both scalability design (planning) and scalability response (operations).

This work applies scalability to both 'Defence' and 'the ADF', though it distinguishes the two. For the purpose of this paper, 'Defence' is considered the broader term, encompassing corporate and strategic aspects of the enterprise as a whole, whereas 'the ADF' is considered more narrowly to refer to specifically military aspects, nested within Defence.

While the focus for this work is Defence and within it the ADF, the principles developed and discussed are intended to be more generally applicable to any organisation seeking to scale.

What Is Scalability?

Scalability is both a property of and a process within an organisation. In the first (design) mode, it is a statement of potential: *how scalable Capability X is*. In the second (response) mode, it describes how an organisation's performance responds to significant changes in quantity or type of workload. Table 1 illustrates these two modes of scalability, mapped to a military analogy.

Table 1. The two modes of scalability: design and response

Scalability …	Statement of …	Military analogy
Design	Potential	*Planning*
Response	Performance	*Operations*

Scalability is best understood with a simple example. Introduced here as a benchmark, Figure 1 presents a classic scaling example drawn from Australia's experience in the First and Second World Wars. As Part 2 will explain, contemporary scalability involves much more than scaling the workforce, and is more complex than scaling a single capability (e.g. soldiers).

Figure 1. A classic example of scaling

Australian Military Mobilisation, World War I and World War II

In the First and Second World Wars, a key Australian military contribution to the Allied war effort was troops, achieved through mobilisation.[2]

From a small permanent professional force and a large part-time militia, the Australian contribution scaled:

- in World War I to the 1st AIF of 416,809 soldiers[3]
- in World War II to almost a million sailors, soldiers and aviators.[4]

Demobilisation occurred immediately following each conflict, rapidly descaling these forces to peacetime levels.

Why Is Scalability Important?

Scalability in Public Organisations

Of all the arms of government, the military enterprise (Defence) has the most compelling need for scalability. Thinking specifically about the Australian Army, the history of the organisation reflects periods of expansion and contraction,[5] the former associated with major conflicts (World War I, World War II, Korea, Vietnam) and the latter associated with the democratic imperative to harvest 'peace dividends', reducing military expenditure during periods of peace. Unlike many militaries globally, the ADF's remit does not explicitly include nation-building duties or a domestic role; nor has Australia historically faced an enduring existential threat justifying a large standing Army. Expansions have generally been rushed and rapid; and contractions have been painful—often bitter and sometimes resisted.[6] ADF doctrine has been limited to a single-paragraph description, without further elaboration:

> *Expansion is the process whereby current force structure is increased in either size (i.e. more of the same) or scope (i.e. new capabilities), and, either way, by appropriate increases to all FIC* [fundamental inputs to capability] *elements.*[7]

For an organisation which has a macro-level purpose (within a liberal Western democracy) to expand and contract to meet demand for warfighting services, the absence of systematic study of scalability as a desirable attribute is striking. This paper aims to correct this omission.

The importance of scalability is better understood when the consequences of omission are considered. In the absence of a conceptual and structured understanding of scalability, scaling events within the life of an organisation risk being unanticipated in the planning, at least initially chaotic (in reality, in perception, or both) in the response, and traumatic in the aftermath. Many Australians may relate to the nation's COVID-19 pandemic experience in these terms.

Cultural analysis provides four important insights as to why scalability—as an organisational concept—has been neglected. First, in the ADF context, the civil–military interface[8] in modern Australia's liberal democracy is at the fringe, rather than the mainstream, of Australian society. 'War' and warfighting are considered exceptional events which are a disturbance to the nation's regular economic and social activity.[9] Second, in an economy structured according to free-market liberal ideals, the unquestioned pursuit of efficiency as an organisational outcome means that scalability design aspects—which may build in ab initio redundancy, precursive to a future expansion—may lack support and funding. Uncritically examined, 'efficiency' privileges short-term outcomes, which disincentivises longer-term organisational planning. Third, there is the uniquely Australian cultural attribute of 'she'll be right, mate'—a tendency to defer action on important initiatives until the last safe moment.[10] Fourth, has been Australia's historical reliance on 'great and powerful friends', and a preference to underwrite national defence through alliance arrangements.

While these cultural insights account for the omission of study of scalability, they also provide the keys to redressing this—achieving a scalability mindset may involve confronting some of the cultural obstacles preventing previous consideration. Cultural obstacles may exist at a macro (national character) or meso (organisational) level.

Scalability in Business Organisations

As business is a beneficiary of a scalability concept, the failure of business literature to tackle the topic is also curious. Literature on lean start-ups and entrepreneurialism tends to focus externally: first on how to scale capture of discrete market segments[11] and create demand for an innovative product; and second on how to raise capital.[12] The latter is raised initially from angel investors for start-ups and then from venture capitalists to scale development of proven concepts and avoid the ‘valley of death’ that exists between an innovative concept and sustainable market share. This literature is less rich on the internal questions of how to scale to meet stimulated demand once funding has been secured. This paper thus also aims to fill this gap in the business literature.

More positively, there are clear advantages of developing scalability theory and practice:

- From a public enterprise perspective, understanding scalability can help in the design and implementation of important public policy initiatives and programs, where delivery is often sought under time and funding pressure.
- From a business perspective, understanding scalability can help achieve business objectives and sustain competitive advantage, which includes exploiting business opportunities (during upscaling), and containing costs (during downscaling).

Consciously understanding the *type* of scaling event an organisation is experiencing is (to paraphrase Clausewitz)[13] an obvious prerequisite for its more successful conduct.

Introduction: Quick Take-Outs

What is scalability?

- How well the performance of your business process, system or organisation responds to significant changes in workload.

Why is scalability important?

- No established conceptual basis, theory or practitioner guidance yet exists in either public or private organisations that are likely to have a scaling requirement.

Part 1: Theory

Scalability is about knowing what, when and how to scale—in multiple directions.

Part 1 scopes the concept of scalability as a first step towards generating a shared understanding and lexicon that can be used as a basis for doctrine development within the ADF. It is presented in three chapters:

- Chapter 1 deep-dives into the question '*What is scalability?*'. It surveys the scalability literature, analysing the range of definitions offered across disciplines. Based on the analysis, three related concepts of performance, cost and reliability are differentiated. From this, an initial working definition of scalability for the ADF is proposed.
- Chapter 2 examines *how to scale*, integrating the literature to develop a structured understanding of scalability in an organisational context.
- Chapter 3 presents an *initial conceptual model* of scalability, designed for the ADF. This conceptual model is applied in Part 2 to derive initial scalability implications for the ADF, and to scope future steps to achieve enhanced scalability.

Tent city at Camp Rocky set up for Exercise Talisman Sabre 2019 (Source: Defence image gallery).

Chapter 1: What Is Scalability?

As a term, ‘scalability’ has noticeably increased in usage within business and academic literature in the last several years. Annex A presents nine definitions of scalability, widely drawn from this literature. The definitions range in their levels of precision and can be broadly categorised as either quantitative or qualitative in nature.

Information and communications technology (ICT)—and, underlying this, data science[14]—is the family of disciplines with the most active scalability literature. In these disciplines, scalability concerns are largely focused on increasing constraints in the capacity of hardware, software, signals processing and telecommunication networks. These constraints have arisen due to the ‘data explosion’[15] of web-based services (from individual websites[16] to the ‘Internet of Things’),[17] artificial intelligence (AI), virtual reality (VR), and computational decision sciences.[18] Anyone who has attempted to purchase major concert tickets online knows the frustrations of an overloaded internet ordering system, website crashes, and the drill of repeated site refreshes! Telecommunication network congestion and power grid blackouts during periods of peak electricity demand are further everyday familiar examples where constraints have prevented scaling in response to changing workloads.

While ICT lends a very technical flavour to scalability and is system orientated, the social sciences represent the second (and relatively recent)[19] family of disciplines well represented in the scalability literature. Education and public policy areas offer important contemporary applications of scalability in practice, and they focus on the challenge of scaling complex social dynamics within an organisation, community or population.

A practical and holistic conceptualisation of scalability for the ADF is usefully framed within these two discipline families. Drawing from each, scalability can be recognised as an organisational attribute with *both* technical and sociological aspects: scalability is a socio-technical construct.

Scalability is a socio-technical construct.

Performance

Before proposing an initial working definition of scalability for the ADF, further dissection of the literature definitions is required. Implicit in several scalability definitions is the related concept of performance; however, it is critical to differentiate these two concepts.

Performance has been defined as the 'amount of work accomplished compared to the time and resources used. Hence, good performance is nothing more than the optimum utilisation of all resources involved'.[20] For an ICT system specifically, performance has been defined as *mean throughput* (number of completed sub-tasks per unit time). The concept of performance can be differentiated from scalability as follows: 'performance measures how fast and efficiently a software system can complete certain computing tasks, while scalability measures the trend of performance with increasing load'.[21]

Based on these considerations, performance is best understood as a *static measure*, whereas **scalability examines how performance changes with changes in demands**—a dynamic, marginalist measure. This distinction is important in designing for scalability, with two significant contributions[22] noting that industry pilot projects typically (and incorrectly) focus testing on performance (or static effectiveness), rather than scalability. Understanding the relationship between the concepts of performance and scalability implies that separate measurement of *both* is a more comprehensive approach, and this insight will be used in Chapter 3 to develop an initial conceptual model of scalability for the ADF.

Cost

The scalability literature also relates **cost** to scalability as follows:

> *Scalable systems can increase or decrease their size with costs that are proportionate to the resulting change in performance. These costs can be monetary or related to other factors such as integration effort,*

> *operator training, or infrastructure upgrades. The options to increase in size to meet growing demand and decrease in size to minimise costs while servicing low demand make scalable systems attractive for completing tasks under uncertainty.*[23]

The relevant inference for the ADF is that cost-effectiveness is a desirable criterion when considering options for a scaling solution. This is considered further in Chapter 2, 'How to Scale'.

Reliability

Reliability is the third concept related to scalability. In order to build a conceptual understanding of scalability for the ADF, it is helpful to distinguish this concept precisely. From the ICT (internet services) literature, '*performance*' includes two sub-metrics: response time and error rate. The frequency of unplanned system outages constitutes '*reliability*'. Dhall[24] stresses that reliability should include 'fault tolerance'—i.e. failure of one part of the system should not result in a complete system failure. Within ICT industries, this is known as 'degrading gracefully'. Contextualising this for the ADF, reliability is likely to be a strongly weighted criteria in assessing scaling options.

Three concepts related to scalability have been presented above: performance, cost and reliability. This list is not exhaustive and may vary depending on the system (or organisation) under study. These related concepts are important system-specific characteristics, useful to distinguish because of their prospective utility both in assessing scaling options, and as metrics to measure scalability effects.

Initial Working Definition of Scalability for the ADF

Having integrated and analysed these various sources of literature, this paper offers the following initial working definition of scalability for the ADF:

Scalability: the ability of the ADF to deliver acceptable performance (internally) and effects (externally) with fluctuation in existing (or new) demands, given contextual constraints.

This definition of scalability is carefully nuanced to recognise both the technical and sociological aspects of scalability noted in the broader literature. 'Acceptable' is used deliberately, noting that 'acceptable' can be subjective and separately assessed by a range of stakeholders.

The distinction between internal performance and external effects is also deliberate. In high-demand scenarios, it is common to subordinate an organisation's functions to the delivery of external effects—sometimes at the expense of important internal activities (e.g. training). While this reprioritisation may be appropriate to meet a short-duration contingency, the health of the organisation will suffer if this subordination is sustained for too long (AKA cannibalisation). Ideally, scalability must aim to deliver important internal activities (performance) simultaneously with external effects. The performance and effects distinction is also reflected in ADF doctrine for metrics, which separately defines measures of performance (MoP) and measures of effectiveness (MoE).

Use of the term 'fluctuation' (cf. 'increasing') in the definition of scalability is intended to signal that scalability is multi-directional. Motivated by issues including cost containment, wastage reduction, minimisation of idle asset time, and efficient resource allocation, the ability to downscale activities is also important. Organisations (including the ADF) are occasionally challenged by equity holders to downscale; achieving this gracefully is clearly included within a robust ADF scalability remit.

The 'contextual constraints' caveat included in the definition reflects that, ultimately, Defence's resource envelope (in both budget and personnel terms) is set by government. However, other large organisations frequently encounter scaling constraints that lie outside their organisation's boundary, including access to capital, market size, shareholder tolerance, government regulation, and social licence to operate. External constraints are not unique to Defence, and sound understanding of scalability can help identify precisely where a scaling constraint lies—and hence prompt thinking on how to address it, if scaling is necessary.

The next chapter builds on this working definition of scalability, addressing *how to scale*.

Chapter 1: Quick Take-Outs

- Scalability is a socio-technical construct.

What's the difference between performance and scalability?

- 'Performance' measures how fast and efficiently a system can complete a given workload. 'Scalability' measures the trend in performance as workload changes. Performance is about static effectiveness; scalability is about dynamic effectiveness.

How is cost related to scalability?

- Ideally, an organisation should seek to scale (workload) with costs that are proportional to changes in performance. Scalability has its own 'cost curve' within an organisation, which needs to consider this curve in decisions to scale (or not). Relatively large fixed costs (as opposed to variable costs) may alter the scaling calculus, as costs are similar whether a large or a small workload is performed.

Why is reliability important for scalability?

- Organisations should seek to avoid a 'single point of failure' which shuts down their entire organisation when workload changes significantly. 'Graceful degradation' means a fault in one area does not immediately become disabling to the entire enterprise.

Definition of scalability for the ADF

- The ability of the ADF to deliver acceptable performance (internally) and effects (externally) with fluctuation in existing (or new) demands, given contextual constraints.

A Royal Australian Air Force C-27J Spartan, joins aircraft from the United States Air Force, United States Marine Corps, United States Marine Corps, Japan Air Self-Defense Force, French Air and Space Force, Republic of Korea Air Force during Exercise Cope North 24 at Andersen Air Force Base, Guam (Source: Defence image gallery).

Chapter 2: How to Scale

The ICT literature on scaling offers well-developed guidance on how to scale, neatly dividing the options into horizontal and vertical scaling. While this literature presents scalability as a two-dimensional concept,[25] the initial working definition of scalability for the ADF extends it by conceptualising scalability as multi-directional. This multi-directionality is intended to signal that scaling up or down may relate to an organisation's existing activities, products or services (i.e. two-dimensional); or it may relate to a *change in shape* for the organisation, with development of new activities, products or services (i.e. along other dimensions).

Horizontal and Vertical Scaling

Horizontal scaling has been described as the addition of more 'units'[26] and is generally referred to as scaling in or scaling out.[27] In contrast, vertical scaling (scaling up or scaling down) refers to changing or activating system architecture—e.g. a structural redesign, with different control node configuration. Understanding the differences between horizontal and vertical scaling strategies[28] is aided by a comparison of theoretical and empirical results in the ICT scalability literature.

A simple but comprehensive theoretical model considers an individual unit within a scalable system to be in one of three states—solo (S), grupo (G) or fermo (F):

> *These states indicate how the unit is working toward completing the task … A unit can be either working in solitary mode (S), interacting with other units (G), or being unproductive due to congestion on shared resources (F).*[29]

Applying three basic scaling laws shows that, at any given point, the most productive state of an individual unit depends on the number of units, and the interaction parameters.[30] This model explains why horizontal scaling can ultimately reach a productivity limit when interaction overheads (coordination costs) exceed interaction synergies.

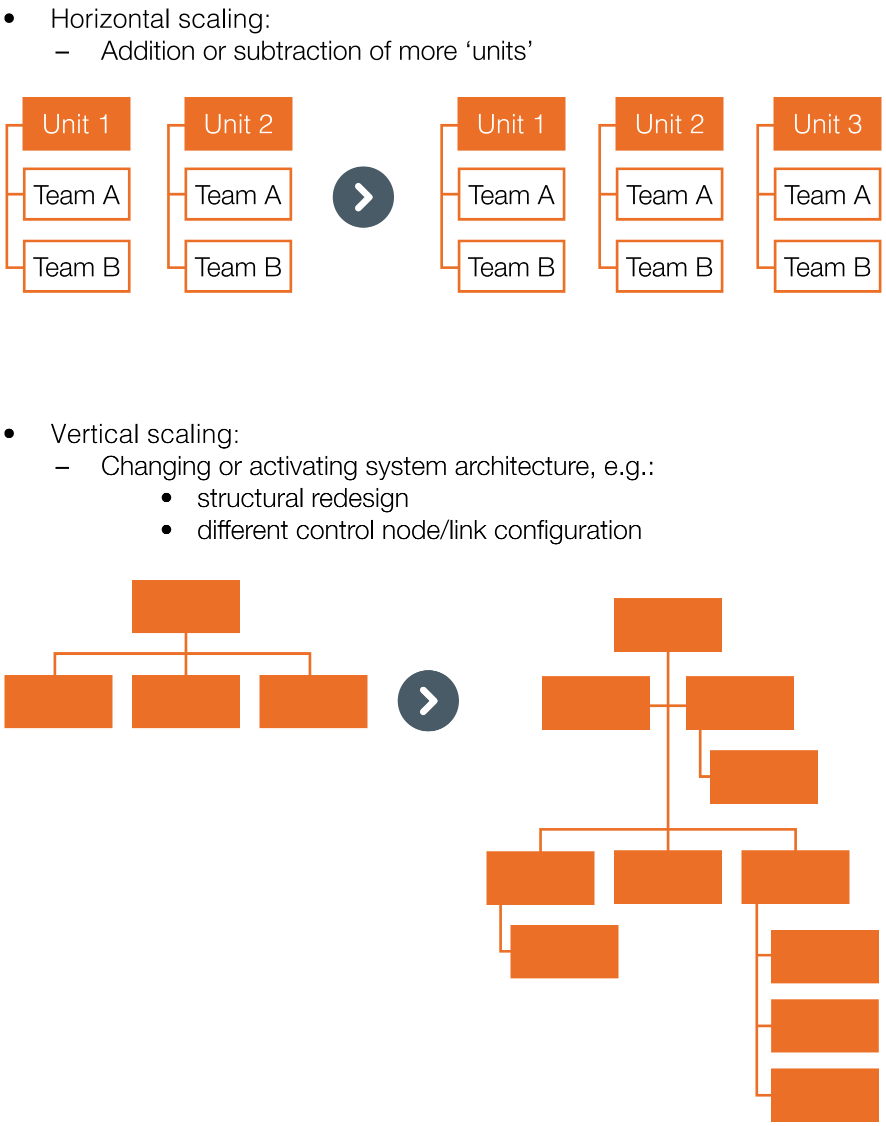
Empirical evidence supports this theoretical finding. In human–robot teaming experiments, three effects on team performance (productivity) were observed with increasing workload (demand):[31]

- **Bottleneck**: where the capacity constraint of a team of fixed size is reached, with further increases in demand remaining unserviced.
- **Saturation point**: at a given demand, adding more units to the team (horizontal scaling) does not increase performance—i.e. the team is performing at optimum efficiency.
- **Degraded performance due to workload**: responding to increasing demand by adding more units actually decreases performance (e.g. due to congestion).

While the bottleneck scenario can be relatively easily addressed by horizontal scaling, the other two scenarios require architectural (i.e. purposeful structural and functional design) solutions—vertical scaling. 'Structural' (i.e. horizontal) scalability is described as 'the ability of a system to expand along a given dimension without drastically changing the system architecture'.[32] Horizontal scalability is also described as 'Type I', with escalation to a vertical, architectural solution as 'Type II', noting that for software developers, poor design that necessitates 'major architectural operations … should be avoided from the beginning at any cost'.[33] A strong preference for horizontal scaling solutions is commonly expressed in the ICT literature, essentially seeking to fully exploit parallel processing. One set of website design rules lists 'Design to Scale Out Horizontally', or out-scaling, as the least costly, most preferable approach, in contrast to upscaling (with the higher cost of more complex system architecture).[34] This conceptual approach has attracted support.[35] The Phorest online platform is an example: as the client base grew, Phorest used 'load balancing' to activate additional servers and redirect new traffic to these servers.[36]

Figure 2 illustrates how to scale, using horizontal and vertical methods.

Figure 2. How to scale: horizontal and vertical methods of scaling. While initially easier, beyond a certain point of horizontal scaling, the coordination overheads exceed the benefits of adding more units laterally, generally necessitating an architectural (vertical scaling) solution.



- Horizontal scaling:
  - Addition or subtraction of more ‘units’

- Vertical scaling:
  - Changing or activating system architecture, e.g.:
    - structural redesign
    - different control node/link configuration

Pushing the Knowledge Frontier—State Transitions to Achieve Scalability

Noting the publication dates of the literature cited above, and the relative recency of key contributions,[37] the knowledge frontier of the ICT scalability literature is characterised by the dawning realisation that there are limits to horizontal out-scaling.[38]

Scalability challenges represent a serious threat to the business model of cryptocurrencies and other cloud-based services founded on middleware,[39] and remain an area of active research[40] in these and other ICT applications, including virtual reality.[41] The scalability challenges of blockchain used by cryptocurrencies such as bitcoin have been described in several 'layers', highlighting that it may be difficult to optimise important system characteristics—decentralisation, security and scalability—simultaneously.[42] In the discipline of AI, the three important characteristics are described as scalability, performance and reliability.[43] An insight from both these applied research areas is that scalability cannot be sought in isolation, and scaling must aim to preserve (or improve) system characteristics deemed important. Extending this insight further, measuring scalability should therefore also include metrics for important system-specific characteristics.

Future research must grapple next with how to achieve effective vertical scaling.

The critical insights from this section on *how to scale* include:

- The initial architectural design of a system influences its ability to scale.
- Ultimately the scalability limit of a given system will be reached.
- It can be postulated that transcending the performance thresholds that are (inevitably) reached with increasing demand involves transitioning the system under study to a different **state**.

A telecommunications application seeking to optimise service under fluctuating demand introduces the notion of *scaling policy*, which defines thresholds (or triggers) for switching between **scalable states**.[44] This implies there is centralised knowledge (monitoring) of the system to be scaled, and metrics in place to measure system state, in order to know when to activate scaling policy.

From Technical to Social

While state transitions and scaling policy are conceptually easy to grasp in technical, mechanistic systems (e.g. ICT), the analogy is more challenging when the scalability target is an organisation (e.g. the ADF), and the more holistic concept of scalability as a socio-technical construct is considered. Here, the social science literature on scalability assists.

Across several social science disciplines, there is an emerging consensus that scalability must overcome the limitation of face-to-face, manual activities (e.g. in education and social intervention settings). In education, for example, scalability is used in the context of **not** increasing teacher workloads as student numbers increase:

> *[S]upporting this form of teaching and assessment in a scalable way by not significantly increasing or using additional course-related resources is deemed crucial. In other words, teaching and assessment methods in the course have to support a larger number of students.*[45]

Automation (e.g. online quizzes) is a 'now-typical' means of achieving teaching scalability. Transition from manual to autonomous systems is considered central to scalability in various sectors, from oceanography[46] to agribusiness.[47] This scalability perspective views technology transition as a means to achieve greater efficiencies—scaling the impact and reach relative to the resource inputs—with the performance goal of scalability of *increasing* returns to the scale. Implicit in these case studies is the assumption that performance (effectiveness) can be maintained following transition.

The challenge of major scaling transitions is also recognised in the social science literature. Applications in public policy consider threats to scalability as those factors which can impair the upscaling of favourable pilot results to larger settings. Three such threats have been classified as:[48]

- **Statistical inference**—how strong must the evidence from a pilot be to justify upscaling?
- **Participant representativeness**—how reflective is the pilot pool, relative to the broader organisation or population?
- **Situation representativeness**—how reflective is the pilot's operating environment of the broader organisation or population?

State transitions are therefore the common thread in the scalability challenges of both the quantitative and qualitative disciplines, with technology representing an enabler. Emerging from this analysis, a refined statement of the scalability goal can be:

Scalability goal: achieving scale through state transitions, while preserving (or improving) other important system-specific characteristics (e.g. performance, cost, reliability, security).

Organisational Context

While much of the literature features the technical enablers of scaling, scalability has been described as both a science (requiring technical expertise) and an art ('knowing the business').[49] Scaling within an organisational context requires more than a purely technical, mechanistic perspective on scalability, recognising the human dimension in organisational endeavours. For example, program scalability has been found to be influenced by factors including leadership, maintenance of relationships, policy windows, financial resources, and political promises.[50] While nominal public policy may aim to scale pilot programs based predominantly on promising initial evidence, this assessment realistically reflects the internal and external constraints many organisations operate within. In a health policy context, for example, it has been observed that the multi-domain nature of scalability requires 'considerable time and knowledge of the service'[51] to be successful. In the ADF context, experience suggests that scaling specific initiatives involves a strong organisational cultural component,[52] especially as the budget allocation may be zero sum (either within or between services). The reception of a new initiative by parts of the existing organisation that are not involved in (or beneficiaries of) the initiative is a significant internal consideration.

Recognising scalability within an organisational context as a socio-technical construct, the key insight for the ADF is that achieving scalability is likely to require *both* technical and human enablers.

Achieving the latter may inter alia require strong attention to education and communication campaigns targeting key stakeholders (internal and external).

The goal of scalability is stated above as ‘achieving scale through state transitions, while preserving (or improving) other important system-specific characteristics (e.g. performance, cost, reliability, security)’. To this set of important system-specific characteristics, ‘culture’ can be added, in two senses. First, a *scalability mindset* is essential to designing organisations to scale. Second, leverage and preservation of positive organisational culture through a scaling response—and especially through challenging state transitions which may significantly change the shape of the organisation—rounds out the macro-level goal of scalability. Scaling can be a transformational opportunity.

Collectively, Chapters 1 and 2 have developed a conceptual understanding of scalability. Chapter 3 applies these trans-disciplinary insights to present an initial conceptual model of scalability for the ADF and considers the implications.

Chapter 2: Quick Take-Outs

- The two basic methods of scaling are **horizontal** (adding more ‘units’, laterally) and **vertical** (changing or activating system architecture).
- Generally, **horizontal scaling** is easier. However, beyond a certain point, the coordination overheads (e.g. congestion) exceed the extra performance of out-scaling. Think of a crowded headquarters with too many subordinate units—the HQ loses responsiveness!
- **Vertical scaling** may involve ‘twinning’, where a partially formed ‘bud’ splits off to create a new branch of the organisation.

Key insights:

- The initial architectural design of a system influences its ability to scale.
- Ultimately the scalability limit of a given system will be reached.
- It can be postulated that transcending the performance thresholds that are (inevitably) reached with increasing demand involves transitioning the system under study to a different **state**.

Scalability goal:

- Achieving scale through state transitions, while preserving (or improving) other important system-specific characteristics (e.g. performance, cost, reliability, security … and culture!), is the scalability goal.
- Recognising scalability within an organisational context as a socio-technical construct, the key insight for the ADF is that achieving scalability is likely to require both technical and human enablers.
- A **scalability mindset** is essential to designing organisations to scale.

Midshipmen and Officer Cadets conduct a march past at the 2023 Australian Defence Force Academy Chief of the Defence Force Parade (Source: Defence image gallery).

Chapter 3: Initial Conceptual Model

For ease of comprehension, this chapter visualises a mechanistic, physical system first—more complex human organisations are considered subsequently. A point about terminology: this paper distinguishes ‘systems’ in a mechanistic sense from ‘organisations’, which includes both systems and human behavioural dynamics.

First-Order Scalability

We start with two premises:

1. Conceptually, scalability is a function of system performance with changing workload.
2. Workload may change in *quantity* (more, or less, of the same) or *type* (existing products and services,[53] or new ones).

Premise 1 is examined here and can be illustrated graphically. Figure 3 shows three basic scalability functions. Assume that the ‘system’ this curve describes is in a fixed state and has not been augmented. This paper defines this as **first-order scalability**. The concave (sub-linear) curve is the most common in real life and may be diminishing (asymptotic, as shown in Figure 3) or—the most common database practitioner experience—declining (e.g. a hump shape).[54]

Second-Order Scalability

Second-order scalability (depicted conceptually in Figure 4) occurs when constraints on system performance are approached and an augmentation (rescaling) of the system occurs. This is a state transition. In a practical, business organisational context, this has been described as ‘crossing the scalability chasm' by 'changing gears'.[55] Figure 4 shows how a system rescaling can initially deliver a step-change improvement in the overall system performance. However—contingent on the properties of the State 2 system—as workload further increases, performance may again diminish. Annex B presents a simple urban road network example of second-order scalability. It is notable that state transitions may involve either horizontal scaling (e.g. widening the carriageway for additional lanes, in the Annex B example) or vertical scaling—altering the system design. Annex C develops a simple ADF scaling analogy: command and control systems and structures.

The following sections extend this conceptual model of scalability by linking with two business literature concepts: value creation, and the theory of constraints.

Figure 3. First-order scalability

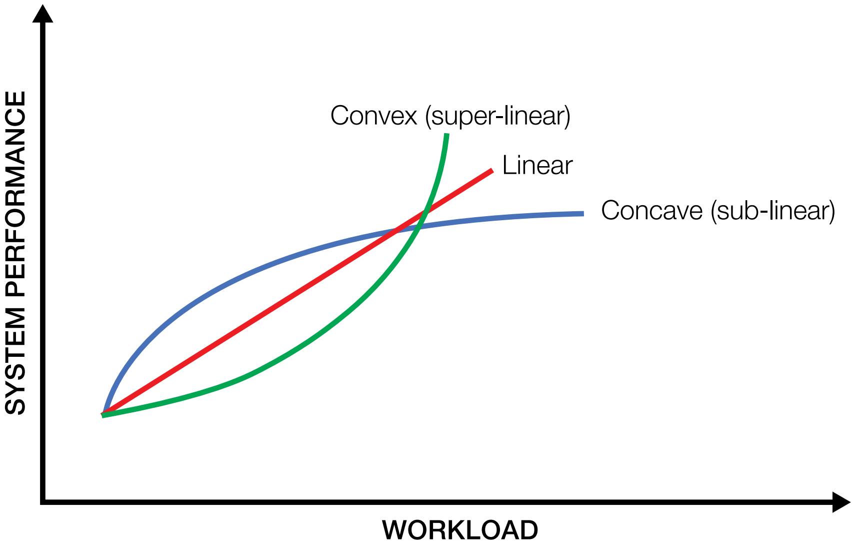
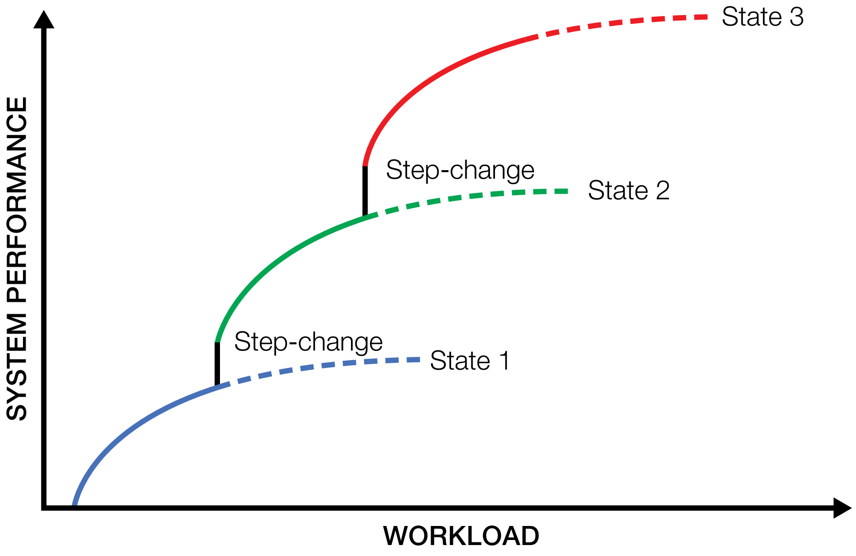


Figure 4. Second-order scalability



Scalability as an Amplified Value Creation Process

Figure 5 represents the value creation process (VCP) of an organisation. Modern variants of the Cobb-Douglas production function[56] hold that outputs are a power function of inputs, with inputs comprising capital (financial resources), labour (human resources) and technology, alongside physical resource inputs (e.g. raw materials). The central circle of Figure 5 is the internal conversion process, where an organisation applies its socio-economic-technical capital to convert the raw input materials to outputs. The VCP can be considered at an enterprise level, or at the level of an individual capability. The more valuable the outputs, relative to the inputs, the higher the VCP.

Shape

This chapter opened with two first-order scalability premises; this section now examines the second premise:

Workload may change in *quantity* (more, or less, of the same) or *type* (existing products or services, or new ones).

If our VCP is considered State$_0$, Figure 6 depicts how an enterprise may respond to a scaling imperative. In this case, State$_1$ represents an expanded VCP, which has also changed *shape* in some dimensions more than others. A scaling event that results in an enterprise scaling proportionally can be considered as isometric scaling, whereas a scaling event that results in some capabilities (for example) scaling more than others can be considered as anisometric scaling. The latter is Premise 2 and is expressed as the scaling ratio parameter (see next section), which captures changes in organisational shape arising from a scaling response.

An organisation that is responding to changes in the external operating environment (cf. simply increased demand on existing services) is more likely to undergo anisometric scaling. One risk of anisometric scaling is that it occurs in an unbalanced (disproportionate) fashion relative to critical enablers. Figure 6 also enables the time dimension of scalability to be visualised.

Figure 5. A generic enterprise value creation process. The circular arrow symbolises the value creation process inside the organisation, as the organisation's ongoing viability is dependent on the perception that outputs, outcomes and effects delivered or achieved by the organisation are worth more than the inputs to the organisation.

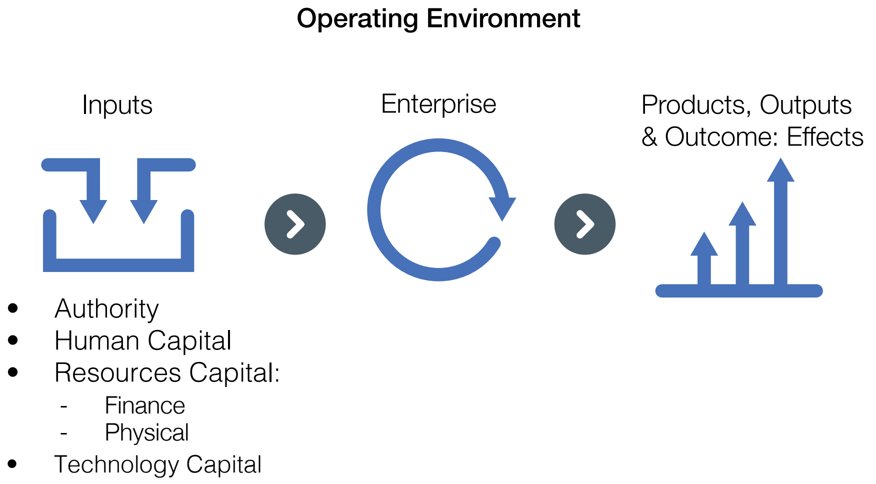
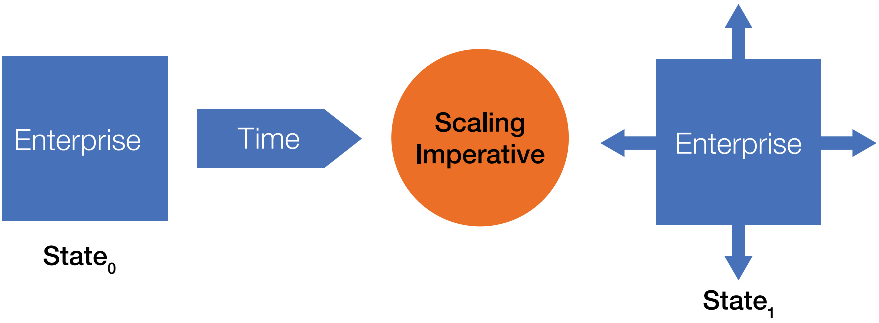


Figure 6. Scaling response to a scaling imperative. While State$_1$ implies that the scaling response is expansion of the enterprise, the scaling response may be a contraction. Also, the shape of the enterprise at State$_1$ is different to that at State$_0$. This symbolises that a scaling response may involve changing an enterprise's shape.



Scaling Parameters

Scalability relates to the speed, efficiency and sustainability with which an organisation can increase or decrease its conversion of *inputs* into delivery *effects* (as experienced by stakeholders). A conceptual model of scalability is not complete without conceptualising a scaling response. There are three essential parameters that are required for a scaling response:

1. **Scaling factor**: e.g. x 2, x 4
   Expressed as a proportion of the start state / original
2. **Scaling rate**: time required to double your throughput, outputs or effects
   How fast to achieve the scaling effect?
3. **Scaling ratio**: shape of the scaled outcome
   Core to enabling

A scaling response, such as that depicted in Figure 6, can be planned based on these three parameters.

This VCP depiction allows scalability to be understood as an *amplified* VCP. This conceptualisation aids initial thinking about scalability because it forces specification of what value a scaling event is creating, and the process by which it is created. It also prompts questions such as: can a product or service of equivalent complexity or quality be produced by another organisation?

Chapter 4, ‘Scoping and Framing’, will propose that contemporary capabilities deliver performance via a *networked* effect. While this effect relates to the product, output or outcome of the VCP, it also applies to the input side of the VCP. Producing a contemporary capability involves complex inputs, and therefore scaling this capability also involves scaling all of its enablers.

The deduction from this analysis is that interdependency mapping of the VCP is critical to successful scaling.

Thinking about scalability as an amplified VCP also prompts consideration of the reverse: value-destruction processes[57] and the diversion of resources away from other activities implied by an amplification of an existing VCP. This diversion incurs an opportunity cost to those other activities, which requires consideration by sponsors in a scaling calculus.

Conceptualising scalability as an amplified VCP is generically applicable to either a public or a business organisation. The unique differentiator of public organisations (e.g. Defence) is the authority and accountability requirements, included in Figure 5, and the linkage to *public* value, explored further in Figure 7.

Figure 7. Public value

Moore’s theory of public value[58] stresses the triangular relationship between:

1. the legitimacy and support conferred by the public to a public organisation (its social licence to operate)
2. the operational capability and capacity developed by that public organisation
3. the **public value** generated by that public organisation.

Collectively, this places a strong obligation and responsibility on Defence to conceptualise and invest in scalability as a *public* value creation process, delivered in the public interest. It is also reasonable to expect that a major scalability event will attract additional public interest and may place additional pressures on the public. Defence must be prepared to manage these challenges, mindful of the social licence conferred by the Australian public, without which Defence cannot operate. This insight is offered as a boundary condition on ADF scalability, consistent with ADF ethics doctrine.

Scalability and Complexity

The above VCP analysis offers a key insight: the more valuable the outputs, relative to the inputs, the higher the VCP. Two additional findings follow from this:

1. The higher the VCP, the more complex this internal process is likely to be.
2. More complex processes are generally more difficult to replicate.

The term *replicability* warrants a brief discussion here, as it relates to scalability. In defining the meaning of a profession, Huntington[59] differentiated a 'professional' from an 'artist' in terms of the former possessing practitioner skill and expertise (a body of knowledge) that can be codified and transmitted to others via education and training. In contrast, Huntington implied that much of the value that an artist creates is the result of natural talent, which (notwithstanding learning basic techniques such as drawing and painting) *is difficult to scale*. The implication of Huntington's analysis is that the replicability of a profession enables it to scale. From this it can be deduced that skills, processes and outputs which are difficult to replicate will also be difficult to scale. Replicability may be particularly challenging if critical inputs are scarce or if the VCP is complex. For example, 'critical inputs' may include items with vulnerable supply chains, and a 'complex VCP' may include an advanced technology manufacturing process rather than a simpler production line.

The basic deductions from this analysis are: the simpler the VCP, the easier it is to replicate—the easier it will be to scale. By extension, high VCPs are more challenging to scale.

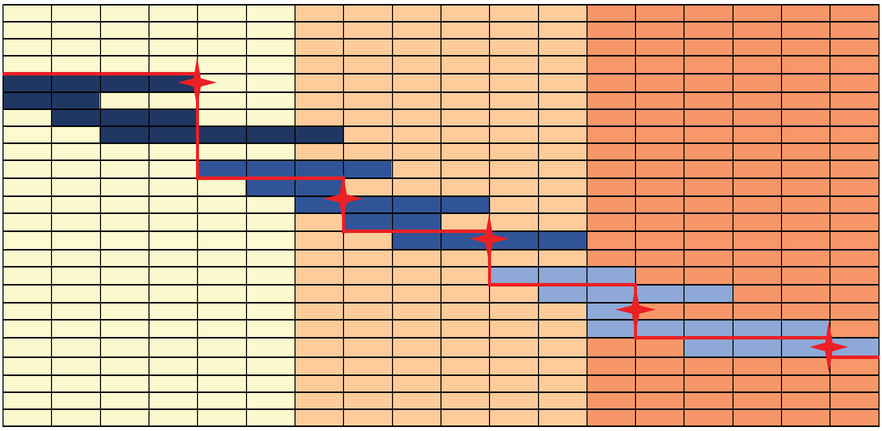
The implications of these deductions for the ADF are explored further in Chapter 6.

Theory of Constraints

Scalability is a particular application of business process improvement (BPI). Understood in these terms, enhancing an organisation's scalability involves identifying and remediating constraints. The following analysis reviews the second critical insight from the business literature, the theory of constraints (ToC).[60] Two elements of the ToC are especially relevant to scalability:

1. At any given point, every organisation has a single constraint that limits operations or performance—the weakest link. If the goal is to increase scale, investment in improving *anything other* than the binding constraint will not increase scale.[61] The ToC prioritises remedial activities towards the binding constraint.
2. Once the constraint is remedied, the organisation is then constrained by some other factor elsewhere in the organisation (and ultimately outside it). In short, every organisation **always has a constraint**, and the scalability leadership task is to 'find and fix' the sequence of binding constraints (Figure 8).

Figure 8. 'Find and fix' the sequence of binding constraints. While you may have many parallel processes occurring across your organisation during a scaling response, there is only one binding constraint at any given point in time (represented by the red stars), and the sequence of binding constraints (represented by the critical path red line) requires your leadership attention to maintain positive momentum during a scaling response.



Business Process Accretion

A third insight is gleaned through the author's organisational experience within the Defence portfolio, and consultation[62] has not yet identified a label in the business literature which captures this insight. Therefore, it is presented in this paper as a novel hypothesis, termed **business process accretion** (BPA), which is described below.

Business processes in old, large organisations tend to accrete over time, as a range of incidents are experienced and responded to with BPIs. While each BPI may be individually well intentioned, over time, accretion can produce an overall business process which is unwieldy and time-consuming: neither agile nor scalable. The administrative and approval process involved in selecting and training new ADF members is one example.

When a BPA is identified, several consequences follow. First, BPAs are prime candidates as constraints on scalability: if a given BPA is regarded as impairing 'business as usual' (BAU), it will almost certainly impair an attempt to scale. Second, BPAs are opportunities for process reform—and in the interests of expedience, it may be necessary for leadership to reconsider the risk profile of the overall process and accept more risk. Third, BPAs may pertain to an overall process that extends across multiple parts of the organisation, and therefore may lack a single overall 'owner'. Here leadership is required to assign clear responsibility and accountability for the overall process; note that this may be challenging given existing structural constraints within the organisation. Finally, BPA is an example of an *internal* constraint—a first-order scalability issue. However, the constraint may be a second-order scalability issue, external to the organisation. Understanding the difference between these two, and activating the appropriate resolution pathways, involves sense-making and influence. These are explored further in Chapter 6, 'Scaling Principles and Their Application to the ADF'.

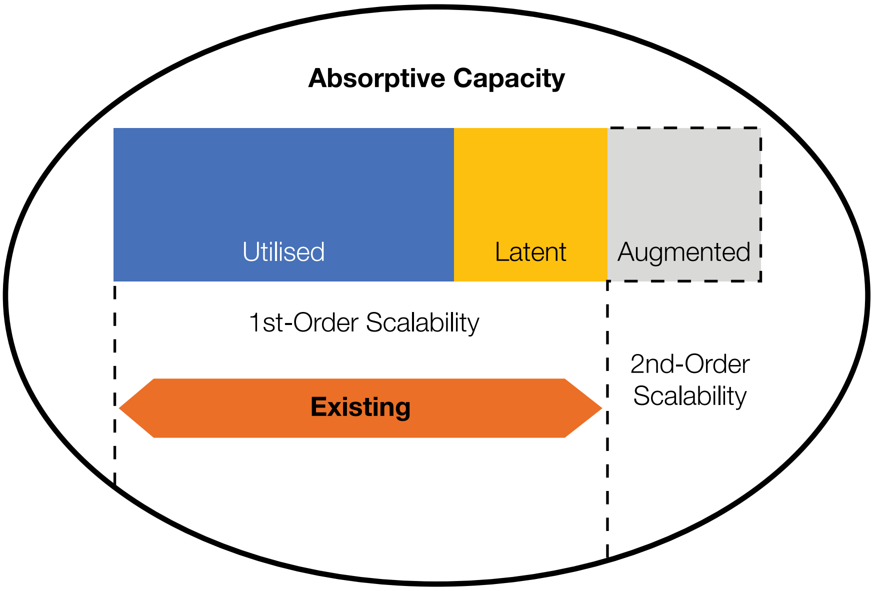
Conceptualising Capacity

Scalability as an amplified VCP involves both the internal business processes of an organisation, and its interactions with its external operating environment—recognisable distinctions in a military context.

The ToC exhorts scalability leaders to 'find and fix' each of the succession of binding constraints that may otherwise stall a scaling response, and highlights that the binding constraint, if not initially, may eventually lie outside the organisation.

This insight prompts a link between the scalability conceptual model and the business concept of *capacity*. Figure 9 conceptualises the business components of capacity, and relates them to first- and second-order scalability.

Figure 9. Conceptualising capacity. Existing capacity consists of two components: utilised, and latent. Existing capacity can be augmented by additional resources. Utilisation of the latent component of existing capacity constitutes first-order scalability. Significant augmentation of existing capacity constitutes second-order scalability. 'Absorptive capacity' refers to the ability of the external environment (or market) to either supply the required inputs for an organisation's scaling, or take up the products, outputs or outcomes of the organisation's scaled effort.



A key deduction from Figure 9 is that the overall scaling response (specifically the state transition from first- to second-order scalability) is likely to be constrained by the absorptive capacity in the external operating environment. Put simply, an organisation's scaling does not occur in a vacuum.

Differentiating Scalability

So far, this work has dealt mostly with what scalability *is*. The final necessary component of a conceptual model of scalability is to differentiate scalability from two adjacent and related concepts: mobilisation and growth.

Scalability versus Mobilisation

Previous (2013) doctrine defined mobilisation as: 'the process of moving from the prepared state for a range of contingencies to being ready to execute a specific operation'. This state is described as:

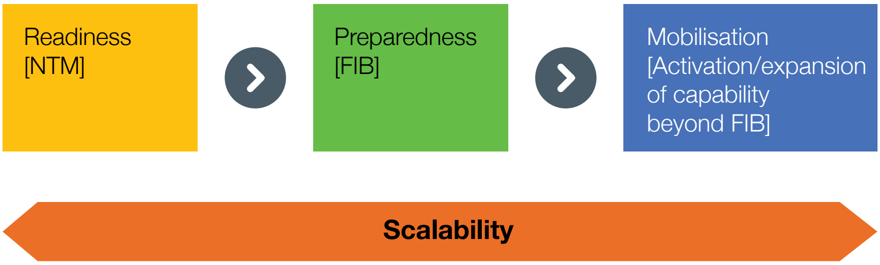
a graduated response across four stages:

- Stage 1 Selective Defence mobilisation
- Stage 2 Partial Defence mobilisation
- Stage 3 Defence mobilisation
- Stage 4 National mobilisation.[63]

The earlier stages of mobilisation (as defined above) do not necessarily imply scaling, but rather internal, routine activities directed towards meeting a specific contingency. More recently, Defence has defined *strategic* mobilisation as ‘the process that generates military capabilities and marshals national resources to defend the nation and its interests’.[64] This definition approximates Stages 3 and 4 of the 2013 doctrine above, and certainly implies scaling. More specifically, strategic mobilisation is a case of upscaling with recourse to resources beyond those ordinarily allocated to Defence, sourced from the national support base (NSB).

Strategic mobilisation is thus differentiated as one specific instantiation of the more general concept of scalability; the latter includes the full range of scaling events at and below the (extreme) threshold of strategic mobilisation. Mobilisation concerns a temporary set of measures for a temporary response, whereas scalability implies standing properties and processes of an organisation that can be leveraged at any time—in short, design principles which are built into an organisation.

Figure 10. The doctrinal relationship between concepts of readiness, preparedness and mobilisation. NTM = notice to move, FIB = force in being. Scalability relates to ease of movement across this spectrum.


Readiness [NTM]
Preparedness [FIB]
Mobilisation [Activation/expansion of capability beyond FIB]
Scalability

The 2013 doctrine relates the concepts of readiness, preparedness and mobilisation as shown in Figure 10. In one sense, scalability relates to the ease with which, and specific mechanisms by which, the ADF can move up and down this spectrum. Thus, scalability reframes the processes of expansion and contraction as BAU in the life of the organisation. Whereas mobilisation and demobilisation are undertaken in response to exceptional events, scalability is conceived as a normalised, accepted and expected organisational activity.

Scalability versus Growth

Growth is considered as an organic process in the life of an organisation. Growth is distinguished from scalability in two senses: incrementalism and intentionality. Growth is considered to occur incrementally based on the ordinary activities of an organisation. In leadership terms, growth is based on 'setting the conditions'—i.e., through providing founding resources and an initial capability for a VCP. In this sense, while growth can be 'cultivated', it cannot be 'mandated'. Rather, it occurs in settings of favourable internal and external environmental conditions, and is not necessarily controlled in its rate, direction or outcome. A garden is a general analogy for growth, and the Australian economy (e.g. gross domestic product (GDP) as a measure of productive output) is a more specific analogy.

In contrast, scalability involves more intention than ordinary growth and is stronger than a superficial response to generally favourable conditions.

Scaling implies specific direction, in response to specific drivers, and for a specific purpose.

Whereas growth is organic, scaling is purposeful and controlled. Scaling implies benchmarking the 'as is' state of the organisation and its environment, envisioning a desired 'to be'[65] state for the former in relation to the latter, and then implementing the specific measures required to achieve that 'to be' state. Scaling can involve both more radical (cf. incremental) interventions and a step-change in outcomes achieved. With respect to time, scalability invokes expectations concerning the rate of internal change and external outcomes to be achieved and may specifically imply a sense of urgency. In short, scaling has a stronger forcing function—an imperative—than ordinary growth. Scaling implies intentionality beyond pure incrementalism.

Chapter 3: Quick Take-Outs

- Conceptually, scalability is a function of system performance with changing workload.
- Workload may change in *quantity* (more, or less, of the same) or type (existing products or services, or new ones).

1. *Initial conceptual model of scalability:*

- **First-order scalability** is where the performance limits of an existing system (or organisation) are reached as workload increases, and may degrade.
    - A state transition is required to unlock further performance if workload increases further.
- **Second-order scalability** involves system augmentation (rescaling).
    - This rescaling can initially deliver a step-change improvement in the overall system performance. However—contingent on the properties of the State 2 system—as workload further increases, performance may again become diminishing.
- Extending this, scalability can be conceptualised as an amplified value creation process (VCP).
- Interdependency mapping of the VCP is critical to successful scaling.
- The simpler the VCP, the easier it is to replicate—the easier it will be to scale. By extension, high VCPs are more challenging to scale.

2. *The three scaling parameters:*

- **Scaling factor**: expressed in multiples of the original state
- **Scaling rate**: time required to double throughput, output or effects delivery
- **Scaling ratio**: the shape of the scaled organisation, relative to the original:
    - Isometric scaling is where the organisation scales proportionally
    - Anisometric scaling is where the organisation scales disproportionally

3. *The theory of constraints (ToC):*

- At any given point, every organisation has a single constraint that is limiting operations or performance: the **binding constraint.**
- If the goal is to increase scale, investment in improving *anything other* than the binding constraint will not increase scale.
    - The ToC prioritises remedial activities towards the binding constraint.
- Once this constraint is remedied, the organisation is then constrained by some other factor, elsewhere in the organisation (and ultimately outside it).
- The ongoing task of **scalability leadership** is to 'find and fix' the binding constraint, and move on to the next constraint.

4. *Business process accretion (BPA):*

- BPA occurs when an initially simple business process grows in complexity over time, in response to various incidents the organisation encounters and then implements control measures for.
- BPAs can produce an overall process which is unwieldy and time-consuming: neither agile nor scalable.
- BPA consequences for scalability include:
    - BPAs are prime candidates as constraints on scalability: if a given BPA is regarded as impairing business as usual (BAU), it will almost certainly impair an attempt to scale.
    - BPAs are opportunities for process reform—and in the interests of expedience, it may be necessary for leadership to reconsider the risk profile of the overall process … and accept more risk.
    - BPAs may pertain to an overall process that extends across multiple parts of the organisation, and therefore may lack a single overall 'owner'. Scalability requires clear leadership alignment of accountabilities and responsibilities.

5. *Composition of capacity in an organisation:*

- First-order scalability:
  - Utilised
  - Latent
- Augmented (second-order scalability)
- Ultimately limited by the absorptive capacity of the external operating environment

6. *Distinguishing scalability:*

From mobilisation	From growth
• Mobilisation is a case of upscaling with recourse to resources beyond those ordinarily allocated to Defence, sourced from the national support base (NSB). • Mobilisation is one specific instance of the more general concept of scalability; scaling includes the full range of events at and below the (extreme) threshold of strategic mobilisation. • Mobilisation and demobilisation are exceptional events; scaling reframes the processes of expansion and contraction as BAU within the life of an organisation. • Scalability *measurement* relates to the ease with which, and specific mechanisms by which, the ADF can move up and down the readiness–preparedness–mobilisation spectrum.	• Growth is distinguished from scalability in two senses: incrementalism and intentionality. • Growth is incremental, based on the ordinary activities of an organisation. • While growth can be 'cultivated', it cannot be 'mandated'. It occurs in favourable internal and external environmental conditions, and is not necessarily controlled in its rate, direction or outcome. • Whereas growth is organic, scaling is purposeful and controlled. • Scaling implies benchmarking the 'as is' state of the organisation and its environment, envisioning a desired 'to be' state for the former in relation to the latter, and then implementing the specific measures required to achieve that 'to be' state. • Scaling can involve both more radical (cf. incremental) interventions and a step-change in outcomes achieved. • Scaling has a stronger forcing function—an imperative—than ordinary growth. • Scaling implies intentionality beyond pure incrementalism.

Part 2: Enterprise Scalability Design and Response

> *If you don't design to scale, you are accepting a finite size of your organisation, and the outcomes and effects it can achieve.*

The introduction to this paper opened with an important premise: that an organisation can enhance its scalability. Part 2 takes up this premise and considers the enterprise design and response aspects of scalability from a practitioner perspective. This topic is covered in three chapters:

- Chapter 4 outlines the scoping and framing considerations and the planning considerations for an organisation requiring a scaling response.
- Chapter 5 presents an industry case study of first-order scalability: the Australian retail supermarket response to the COVID-19 pandemic.
- Chapter 6 derives scaling principles and metrics, then deep-dives on Defence implications in terms of Australian military strategy and capabilities.

Royal Australian Air Force personnel during Exercise Cope North 2024, Andersen Air Force Base, Guam, USA (Source: Defence image gallery).

Chapter 4: Scoping and Framing

Enterprise scalability design and response reflects the intentionality of scalability, as the standing set of properties and processes of an organisation which can be leveraged at any time to achieve a scaling outcome.

Part 1 presented scalability as a socio-technical construct. For the Defence enterprise specifically, the ADF's capstone concept, Concept APEX,[66] underscores the importance of integrating the human, procedural and technical dimensions. Applying this intuition, the first step in both enterprise design and response is scoping and framing[67] scalability, presented below as an analysis of **why**, **what**, **when**, **where** and **who** parameters.

Why?

Understanding why an organisation is scaling involves appreciating both the scaling imperative and the scaling intention.

1. The **scaling imperative** can involve either internal or external drivers. The significance of this difference lies in attracting support and scrutiny for the scaling effort. For example, an externally initiated scaling event (e.g. in response to a government direction to Defence) may attract both additional support and additional scrutiny. However, an internally initiated driver may not attract support within the organisation and may be challenged to win resources from external sponsors unconvinced of the need for or reluctant to commit additional investment.[68]
It may be deduced that the initial source of the scaling imperative will be the strongest advocate for the scaling event.
2. The **scaling intention** involves understanding what outcomes and effects the scaling event is seeking to achieve, and thoroughly exploring the most optimal means of achieving this. At the appropriate level, it is important to analyse and test both the stated problem

and the range of possible solutions. For example, if the problem involves the organisation's supply of goods and/or services to meet an external demand, in a scenario where supply fails to meet demand, two high-level solutions are: (1) upscaling supply and (2) downscaling demand. Empirical observation suggests that (1) is often favoured and (2) neglected, though the latter may represent the more cost-effective solution.[69]

What?

In the Defence context, a critical decision is the selection of what needs to be scaled. There are essentially two schools of thought informing this decision. The first considers that, by studying potential future war scenarios, we can anticipate which specific capabilities will be more useful in the next conflict. The second school espouses that prediction is fraught, and that militaries are better placed to rapidly adapt to each new contingency as it emerges[70]—i.e., to identify which capability is working in a new contingency, and rapidly scale that capability. In practice, the ADF blends elements of both schools.

The introduction to this paper described scalability as involving changes in the *quantity or type* of workload. Identifying what needs to be scaled can be prompted by two bifurcation questions:

1. **Existing capability or new capability?** While scaling existing capability involves quantity, scaling a new capability involves type. Prima facie, it can be reasoned that the latter features the introduction into service (IIS) process. The deduction from this bifurcation question is that (at least conceptually) scaling an existing capability is easier than scaling a new capability, due to the additional complications associated with IIS.
2. **Hardware (assets and infrastructure—the technical dimension of scalability) or software (people and processes—the human and procedural dimensions of scalability)?** In the classical scaling example presented in Figure 1, the majority of Australia's military contributions to World Wars I and II involved soldiers; therefore the scaling challenge was predominantly about people. While separating hardware and software elements eases conceptual understanding, current and future scaling challenges may involve both elements

simultaneously, as modern military effects are produced through integration of assets with personnel. For example, a nuclear-powered submarine capability involves a substantial crewing consideration, in addition to the actual hardware assets. A more sophisticated rendering of this bifurcation question involves specifying the capability holistically, including its various enablers—i.e. the fundamental inputs to capability (FIC). For many contemporary capabilities, their performance is a *networked effect*; therefore by implication, process understanding of interdependencies is required to effectively scale.

From a funding and availability perspective, the technical component of scalability—major platforms and capabilities—can be considered as relatively **fixed elements** within the Defence enterprise, in the sense that their funding lines are locked into long-term contracts that require senior approval to modify, and their production and construction times are often multi-year. By contrast, workforce is an example of a more **variable element** of the Defence enterprise, due to the historical perception that soldiers can be rapidly recruited and trained, and rapidly demobilised following a period of conflict. From a service perspective, Army has the largest workforce of the three services, and has therefore historically been the obvious target when an up- or down-scaling response is required by Australia's military. Whether the perceived high scalability of the human dimension—soldiers as a capability—remains relatively true in the contemporary operating environment is now contestable, given, inter alia, the barrier of high technical acumen and proficiency required of modern soldiers to be combat-effective.

When?

The 'when' scalability parameter involves time, with two sub-parameters: urgency and duration.

1. **Urgency:** *How fast?* The concepts of growth, scalability and mobilisation can be ranked, in that ascending order, by the time urgency of the process. Scaling rate can be both externally prescribed (e.g. a deployment deadline) and internally constrained. Together, these factors place a premium on accelerating internal scaling processes—the procedural dimension of scalability. To achieve scalability under these conditions, organisational leading practice involves applying business process analysis[71] and the ToC,[72]

optimising on the time dimension. This involves documenting each step in a scaling process, and its duration. The overall length of a process can then be examined, and the most time-consuming steps—which constrain the overall scaling rate—identified. The ToC counsels that investment in the most time-consuming step is required to increase the overall scaling rate.
The means to achieve this may include automating previously manual steps.

2. **Duration**: *How long?* A scaling response may be required for a temporary surge or sustained as a protracted effect. Expedient measures may be sufficient for the former, but may present risks if sustained. For example, internal degradation may occur if important but non-urgent organisational functions—e.g. training—are reprioritised as part of the expediency. If a long-term scaling response is required, more permanent structural response and resourcing may be required.

Where?

The 'where' scalability parameter involves geography, and whether a scaling event is conducted in a centralised or a decentralised manner, both from geographic and from command and control (C2) perspectives. If concurrent responses are required in multiple locations (e.g. the Black Summer bushfires of 2019–2020 and the COVID-19 pandemic),[73] a decentralised scaling response may be required, mobilising local resources. In the Australian context, the state jurisdictions form a natural set of decentralised nodes, able to leverage pre-existing governance and infrastructural networks.

Who?

The 'who' scaling parameter involves mapping stakeholders and their relationships in a scaling process. Stakeholders may include:

1. Sponsors—e.g. internal and external leaders who provide authority, resources and support
2. Targets—e.g. new recruits
3. Enabling personnel—e.g. qualified instructors
4. Commentators—e.g. media and other removed influencers who may offer public comment on or criticism of a scaling process.

Indirect stakeholders include the Australian public, from whom Defence ultimately derives its social licence to operate. The purpose of identifying the range of stakeholders in a scalability context is that a scaling event may activate an expanded stakeholder base, and may also require more active management of Defence's relationships and reputation. This is true whether stakeholders are participating on a voluntary or on a mandatory basis. Maintaining and enhancing goodwill encourages more discretionary effort towards a scaling event. Operations in the information domain are likely to be critical.

Collectively these scoping and framing parameters serve as the initial planning considerations for a specific instance of scaling. They also capture the considerations relevant to an organisational sense-maker seeking to enhance their enterprise scalability design generally. Organisational sense-makers can be described as those individuals who possess knowledge of their own local node, and of the wider network of links and nodes, with their dynamics through time.

Planning Process for a Scaling Response

The ADF is well skilled and practised in both deliberate and expedient planning for responses to emergent contingencies—e.g. using the Joint Military Appreciation Process (JMAP).[74] Derived from the JMAP, this paper proposes the following five-step process where a scaling response is required:

1. Identify the sequence of scalability constraints.
2. Develop scaling options—including consideration of managing both supply and demand (for specific services). If the latter is not possible, the scaling options must address the binding constraint(s) to be successful, in accordance with the ToC.
3. Evaluate scaling options (according to criteria which may include performance, cost, reliability, security, timeliness).
4. Select and implement the preferred option(s).
5. Measure and monitor scaling effects.

Once a scaling response is underway, monitoring the response through application of the ToC may identify three typical effects on performance in response to significantly increased workloads: bottlenecks, saturation points and workload-induced performance degradation (discussed in Chapter 2).

Identifying these separate effects allows the scaling response to be adjusted in stride to address them. In the sense of connecting ends (the intent of the scaling response), ways (methods—the options to address the scaling constraint) and means (harnessing the additional resources required for the scaling response), the above planning process generates a *scaling strategy*.

The purpose of Chapters 1 to 4 is to equip practitioners with some ab initio considerations when embarking on scaling design or response. Chapter 5 presents a case study of a significant *unplanned* scaling response in real time: the Australian retail supermarket industry response to the COVID-19 pandemic. The study of the scaling strategy developed in this scenario provides scalability insights that are relevant for Defence.

Chapter 4: Quick Take-Outs

To scope and frame a scaling response, determine the **five Ws**:

Why? Clarify scaling intention, and scaling imperative (internal or external driver?)

What? Are you scaling:

- an existing or new capability?
- hardware (e.g. assets and infrastructure) or software (people and processes)?

When?

- Urgency: How fast? (Scaling rate)
- Duration: How long? (Expediency or sustainability)

Where? Centralised or decentralised? Control measures?

Who? Who are your:

- Sponsors?
- Targets?
- Enabling personnel?
- Commentators?

To plan a scaling response, start with these five steps:

1. Identify the sequence of scalability constraints.
2. Develop scaling options:
   - Supply-side
   - Demand-side.
3. Evaluate scaling options
   - according to criteria which may include performance, cost, reliability, security, timeliness.
4. Select and implement the preferred option(s).
5. Measure and monitor scaling effects.

Soldiers from Combined Joint Task Force 637.3 hold an Australian flag ahead of an official photograph of the 2023 Pacific Games security forces at Camp Clark, Honiara (Source: Defence image gallery).

Chapter 5: Australian Supermarket Response to the COVID-19 Pandemic—a Case Study in First-Order Scalability

> *The Industry has lifted itself through some glass ceilings of what it thought it could do and can do*[75]

Preamble

As outlined in Part 1, a scaling imperative can be either internally or externally driven. The COVID-19 pandemic represented a significant external scaling imperative for most organisations in Australia. While some organisations had to rapidly downscale their operations to a 'tick-over' mode in response to rolling lockdowns, other organisations—including the retail supermarket chains—were suddenly faced with an upscaling imperative. Demand for certain goods and services increased, alongside a need to provide surety of continued supply as both local and international supply chains buckled and panic-buying spiked.

The retail supermarket industry response to the COVID-19 pandemic represents an opportunistic case study of an external, unexpected scaling imperative. The purpose of this case study is to empirically apply the scalability theory developed in Chapters 1 to 4, and to identify industry scaling insights of potential relevance to Defence.

Prima facie, the retail supermarket industry has several similar characteristics to Defence, including:

1. a high degree of standardisation in operations—well-documented protocols and procedures, and mandatory compliance with standards

2. a high degree of hierarchical leadership, balancing considerable ‘mission command’ granted to store and warehouse managers with absolute primacy of head office chain-of-command
3. a workforce- and logistics-intensive core business, involving complex supply chains and delivery of diverse effects
4. national organisations which nonetheless, during the pandemic, needed to work with Australia’s federated governance system and state jurisdictions with separate COVID-19 rules and protocols
5. the requirement to deal with the challenge of Australia’s expansive geography, and widely dispersed population.

In Australia, the retail supermarket industry is dominated by the ‘Big Four’ chains: Coles, Woolworths, Aldi and IGA. This case study is based on interviews with two senior supply chain managers, one from Coles and one from Woolworths, separately in June 2022. While not encompassing the entirety of the industry, together Coles and Woolworths do hold the majority of the Australian market share and are therefore considered representative of the retail supermarket industry.

The interviews were semi-structured and were undertaken with interviewee consent and on the condition of anonymity. The interviews were conducted in a three-part chronological sequence with respect to the retail supermarket industry experience of the pandemic: planning, response and reflection. Table 2 summarises the key observations shared by the interviewees, as well as the initial, tactical scalability insights drawn from them.

Table 2. Summary of scalability observations and insights from the Australian retail supermarket industry response to the COVID-19 pandemic

	Observation	Type	Scalability insight
1	Limited contingency planning in advance: strong reliance on leadership once a contingency arises. The key leadership task is to prioritise	Planning	1. Without explicit planning, a scaling response tends to be limited to first-order scalability (i.e. redirection and fuller utilisation of the organisation's existing personnel, systems, processes, infrastructure)
2	Dispatch of full pallets of (non-perishable) product to stores, rather than JIT 'picking' of pallets optimised to each store's historical sales demand	Process	2. Change procedure 3. Be willing to accept redundancy (cf. efficiency), in this case by dispatching more product than usual (oversupply to stores considered less risky than undersupply, given likelihood of disruption in each link of the supply chain)
3	Operation of stores with fewer people	Realising operational efficiencies	4. Release existing workforce capacity to redirect to higher priorities
4	Reduced offer: • Reduced range in a product category • Reduced shelf stock of product	Performance measures	5. Be prepared to reconsider 'acceptable' performance standards and offer a reduced or adjusted value proposition to customers
5	Acceleration of online delivery	Accelerating implementation of existing service	6. Scale a specific, existing function. A marginal, non-core service offering can suddenly become core

	Observation	Type	Scalability insight
6	Emphasis on self-checkouts	Reshaping existing service offering	7. Weight an existing function and redirect released workforce capacity accordingly
7	Pop-up distribution centres	Innovation	8. A low-tech innovation was used to decentralise distribution nodes closer to stores/customers, to avoid interruptions to the main supply route caused by border closures. An ADF logistic analogy is 'pushing' third-line logistics closer to the deployed units
8.	Reduction in store trading hours, imposition of density limits and product limits	Reshaping existing service offering	9. Implement control measures to even out demand and provide time for the logistics system to catch up
9.	'Christmas surge' protocols enacted as BAU	Process	10. Apply experience from previous surge events
10.	Extended use of technology, automation and data analytics (e.g. barcode scanning) to: • manage customer numbers in store • use transaction data to predict peak periods	Technology as an enabler	11. Leverage pre-existing advantages in automation and technology
11.	Use of Zoom software to communicate and coordinate when border closure prevented movement of people	Redundancy in communications	12. Maintain communication, coordination and control via alternative means

	Observation	Type	Scalability insight
12.	Development of 'confirmed case' management procedures when infection occurred in chilled product distribution	Rapid procedural development	13. Leadership was exercised ahead of government direction to develop protocols to manage isolation and deep cleaning, to maintain business continuity
13.	Existing liaison mechanisms between retail supermarket industry and government were applied and expanded to develop, implement and coordinate protocols across store networks	Stakeholder collaboration	14. Activate and expand liaison architecture, using pre-existing governance frameworks 15. A government–industry partnership was established, with government moving quickly to consult
14.	Centralised control by a small, dedicated leadership team with clear: • accountabilities • delegations • mechanisms to accelerate if necessary	Leadership	16. Emphasise communication, coordination and control given the complexity of the response (in this case across the store and distribution network); requirement for consistency in policies and messaging
15.	Lease of under-utilised warehouse space to support decentralised distribution centres	Process	17. Access latent asset capacity, even if 'inefficient' and not owned
16.	Modified shifts of delivery, warehousing and store staff	Process	18. Access latent workforce capacity
17.	Temporary relaxation of curfew delivery hours negotiated	Stakeholder collaboration	19. Adjust rules to access latent time capacity
18.	Prioritisation of serviceability of essential over discretionary items	Prioritisation	20. Redirect capacity currently utilised on discretionary product lines towards essential products

This case study is structured into three sections. Section 5.1 describes the supermarkets' pandemic experience in the planning–response–reflection sequence, from the perspectives of the two executives. From this, the emergent strategy used by the supermarkets is described. Section 5.2 extends the analysis of this strategy and of Table 2, identifying and discussing the industry's scalability insights at a strategic level for Defence. Section 5.3 applies these insights to Defence's pandemic experience, and summarises the case study.

Retail Supermarket Experience of the Pandemic

Planning

Interviewees described the COVID-19 pandemic experience from an Australian retail supermarket perspective as a 'perfect storm' scalability scenario: simultaneously increased demand (typified by panic buying) and decreased supply (due to supply chain disruptions). Both executives explained that while the industry (and individual supermarket chains) consider high-level corporate risk scenarios, generally these scenarios are more localised and involve either a supply shock or a demand shock—but not both. Neither chain explicitly plans for negative scenarios at the scale of the COVID-19 pandemic. One executive commented: 'We might plan [for scenarios where we] lose a node, *but not three-quarters of all our nodes* … people [generally] only plan for what they believe they are capable of recovering from.'

In terms of scale and concurrency, the COVID-19 pandemic was unprecedented. Therefore, the approach adopted by both supermarket chains was to adapt to the contingency as it evolved.

Response: Emergent Business Continuity Measures

Both interviewees stressed that their organisations are acutely aware that their activities are not just an essential service, but an *emergency service*. In acknowledging this, they espoused a strong sense of corporate social responsibility to meet the COVID-19 challenge and work closely with stakeholders, including government. Both executives stressed the partnership between industry and government, united by the common purpose of maintaining food supply to Australia's widely dispersed population.

One executive stressed the uncertainty caused in both customers and the workforce by the pandemic, and the importance of leadership in instilling confidence that this situation could be safely managed. The other executive noted how frontline staff in their organisation stepped up, realising they were 'serving the nation' and the importance of their role. This boost in morale proved critical in workforce acceptance of changed working conditions (e.g. longer shifts) and of the need to perform extra shifts as a temporary surge response.

From discussion with the interviewees, an overall emergent strategy can be derived for the supermarkets' response to the pandemic. Just as the pandemic's 'shock' to the supermarket system involved both supply– and demand-side features, the emergent strategy also addressed both these sides.

Figure 11. Supermarkets' emergent strategy

The supermarkets' pandemic response strategy largely consisted of:

- on the **supply side**, redirecting current and accessing latent capacity within existing systems, structures, facilities, workforce and resources
- on the **demand side**, communication with stakeholders to adjust expectations for the standard of service (emphasising essential over discretionary items, and quantities).

Internally there was some *downscaling* of non-essential functions, and redirection of effort towards higher priorities (e.g. online delivery). However, the majority of the pandemic response consisted, on the supply side, of innovatively accessing the latent component of existing capacity (Figure 9), often through changing processes or temporarily removing constraints (e.g. curfews on after-hours deliveries to supermarkets, due to resident noise concerns).

One executive described the considerable scope, within warehousing, for new trade-offs during the pandemic. They described one such utilitarian trade-off (based on reduced product range, with the same warehousing staff and resources) as follows: 'We can pick 1,000 cartons for 200 stores … or 1,500 cartons for fewer stores. [Our guiding principle during the pandemic was to service] the bulk of the need for the bulk of the people.' This executive also explained that their chain was able to access spare warehousing space from across its network, even if not owned, and that this opened up additional capacity. While this arrangement would be considered inefficient under BAU conditions, the redundancy became valuable during the pandemic.

The same executive observed that optimising their chain's response to the pandemic placed a significant premium on business intelligence, and specifically the organisational sense-makers who understand the chain's networks at a meta level, not just local area networks. They specifically highlighted the premium on understanding how much latent capacity was in the network and where it could be found (e.g. unused hours on a delivery vehicle; a second driver prepared to work night shift).

Essential versus Discretionary

One executive described the pre-pandemic supermarket–consumer relationship as one in which, through competitive forces, supermarkets strove to meet every possible discretionary consumer preference by offering maximum range, with high and undifferentiated internal standards for shelf stocks. They described how the pandemic forced supermarkets to think in utilitarian terms of essential versus discretionary items, and to prioritise the former at the expense of the latter. In fact, it was the large capacity previously dedicated to discretionary items that was internally redirected to service more essential needs. This required the supermarkets to redefine the supermarket–consumer relationship by communicating extensively with customers, encouraging their adjusted perspective towards a reduced product range and lower shelf-stocking rates.

Demand management was actively engaged in by both chains, and involved measures including:

- reduced opening hours
- strict in-store density limits, monitored by digital check-in
- per-person product limits
- hygiene and infection control measures, pioneered in advance of government mandates in numerous instances, including sanitiser, face masks, COVID marshals, routine cleaning, and confirmed-case deep-cleaning protocols.

Reflection: Legacies and Organisational Learnings

Both executives spoke positively of increased awareness of food security in the Australian Government and the Australian public. One executive stated that the pandemic has triggered several long-term shifts with an enduring impact on their operations. First, internal migration of the Australian population has appreciably redistributed demand (for supermarket services) and supply (of labour for workforce). They described previously modest regional and rural districts now experiencing boosted tree-change/sea-change populations. Second, there is sustained consumer interest in product origin, with a significant proportion of consumers preferring Australian-sourced goods. Third, from a retail supermarket perspective, COVID-19 has prompted deeper thinking on resilience, noting that this concept applies differently to different product lines:

- Fresh food **must** be delivered continuously, on the basis of 'just in time' (JIT) logistics.
- Packaged goods are imported based on established demand patterns, require considerable lead time, and are less responsive to sudden shifts in demand (e.g. for toilet paper). Therefore, resiliency for this product line requires a 'just in case' (JIC) logistic approach, potentially involving more warehousing.

One executive noted indications that consumers were permanently changing their preferences, with a more flexible mindset: 'We are willing to accept something that is different—penne *versus* rigatoni. Everyone has to compromise.' They also highlighted that the pre-pandemic supermarket–consumer relationship had 'forced a gold-plated solution' and that now, during a crisis, 'maybe we can accept bronze'. They described supply chains as returning to a more normalised level, but noted that emerging from the pandemic presented opportunities to:

1. re-examine certain paradigms in standards, and 'look at processes … what's the new normal? Do [stores really need] to have on-shelf availability at 95 per cent?'
2. embed realised efficiencies—for example, stores are now capable of permanently operating with 20 to 30 per cent fewer staff
3. permanently shift the supermarket–consumer relationship, and 'nudge' consumer preferences rather than simply accepting and striving to meet them.

- Finally, one executive reflected that, while the crisis ‘brought people together’, sustaining these heightened levels of cooperation may be challenging. In terms of relationships with government, they highlighted that industry was developing leading-practice protocols for infection control and confirmed-case management. Further, they observed how government accepted and formalised this emergent practice into mandatory COVID-19 requirements more broadly. This executive cited the ongoing importance of consolidating this leadership relationship with government to ensure a quality and practical regulatory environment into the future. In terms of internal business operations, they remained confident that the COVID-19 pandemic experience has built in (through corporate memory) enhanced ability to respond to the same or similar shocks (in either demand or supply) in future.
- Their key leadership lesson is the importance of placing the leader ‘in the best position’ to guide the response. In contrast, one executive highlighted the risk of returning to BAU without consolidating the learnings of the pandemic. They commented that ‘the retail industry was caught short with lack of planners and process improvement opportunities’. They further stressed the need for more education and professionalisation in the sector, alongside documentation of the lessons learned from COVID-19.

Industry Scalability Insights for Defence

As described in Part 1, the main distinction between first- and second-order scalability is that the former essentially involves either redirected or full(-er, -est) utilisation of existing capacity and resources, whereas the latter involves a state transition that significantly augments capacity. Analysis of observations captured in Table 2 suggests that the Australian retail supermarket industry response to the COVID-19 pandemic was largely a first-order scaling event.

In systems theory terms, the COVID-19 pandemic was an example of an external scaling imperative in the form of a ‘shock’. An ideal ‘resilient’ system does more than recover from a shock—it learns and builds muscle memory which increases its coping mechanisms to face the same or similar shocks in the future. In response to shock, leaders give various amounts of emphasis to the importance of documenting the processes and procedures that arise. The variables are linked to whether an organisation’s leadership believes that the shock experienced was truly exceptional and a one-off, or that it is likely to represent a step-change in the operating environment, necessitating a more permanent response. There is a tension between reliance on operational response and more systematic planning, and the role of leadership across these two distinct functions. This tension is central to the scalability insights for Defence that are drawn from this case study. The top five strategic insights are presented below.

Insight 1: Crisis (within resilience range) will generally initially involve first-order scalability

A critical observation is that the supermarkets did not plan for a contingency at this scale, and they needed to respond immediately and in stride: that is, maintain business continuity while rapidly implementing change from the options already available from existing capacity and resources. The scalability insight drawn from this observation is that the scaling response to a *crisis* (an event which is unplanned, major and requiring immediate response) is more likely to be first-order in the first instance. Whether the response escalates to the more challenging second-order type can be deduced as a function of two crisis factors—magnitude and duration:

1. **Magnitude**. While no doubt *inconvenient*, even in the worst stages of the pandemic, the response offered was still *acceptable* in the sense of the ability of the organisations (supermarkets) and their principal stakeholders (customers) to cope with the revised offer. From a basic food supply perspective, the COVID-19 situation did not become life-threatening to Australians; nor did widescale market failure occur. In short, the magnitude of the COVID-19 pandemic was within the resilience range of the existing supermarket system. Targeted measures to address the needs of the most vulnerable (for example an 'assistance box' of everyday basics for every pensioner, and 'pensioner-only' special opening hours) were critical to maintaining a general public perception that the situation was under control, and assisted with revising perspectives on what constitutes hardship for oneself relative to others.
2. **Duration**. While the intense lockdowns in Australia lasted months (and were not experienced as short term in the view of many Australians), nonetheless this period was *manageable*, from the supermarkets' perspective, with existing capacity and resources,

> and is assessed in retrospect as a temporary surge. A contingency requiring a permanent or far more protracted response would require a state transition (in scalability terms).

The implication for Defence is that, where a response is required which is *outside the resilience range*, an intervention may be required to achieve a second-order scalability response. An intervention may involve the significant commitment of additional inputs—i.e. authorities, time, capital, resources, planning. See the VCP presented in Chapter 3 and Figure 4.

Insight 2: Understanding your organisation's redirectable and latent components of capacity is key to a first-order scaling response

The supermarkets' emergent strategy applied two responses: redirection of capacity and fuller utilisation of latent capacity. A critical insight from this study is that there was, in fact, significant capacity within the supermarket system for potential redirection. Specifically, this capacity consisted of the resources and effort usually invested in servicing a significant proportion of discretionary consumer items. The pandemic forced a simple reprioritisation of supermarket capacity from discretionary to essential items.

In the Defence context, scalability decision-making on redirection requires that judgments are made concerning discretionary versus essential activity. The capacity for immediate redirection is contingent on two factors: the pre-existing proportion of discretionary as opposed to essential activity; and the degree of *fungibility* in the skill sets. In economics, fungibility is the extent to which a given resource is interchangeable or transferrable across different functions. A force with limited initial discretionary activity and with low skill-set fungibility will have limited capacity to redirect—a first-order scaling response. This scenario implies that early recourse to second-order scalability (with significant augmentation) may be required.

The second supermarket response, fuller utilisation of latent capacity, was largely achieved by changing existing rules and procedures—for example, 12-hour rather than 8-hour shifts. In the Defence context, a similar concept is referred to as the level of capability (LOC) spectrum. Within this concept, the gap between the minimum (MLOC) and operational (OLOC) represents potential latent capacity. A unit held in MLOC status has under-utilised capacity. Such a unit can be brought to OLOC with additional resources. MLOC is an efficiency, in the sense that it costs more to hold a higher proportion of the force at a higher directed LOC (DLOC) or OLOC—just as it costs more to pay night-shift rates to staff. A scaling imperative implies that the organisation considers that the extra cost of accessing this latent capacity is warranted.

Scalability decision-making on accessing latent capacity requires the ADF to make at least two judgments:

1. the extent to which BAU rules and procedures can be changed. This may involve risk tolerance (re-)considerations
2. whether the extra access costs (both direct and indirect) are warranted. For part-time personnel, these costs include the indirect opportunity costs to the broader economy (e.g. Reservists will generally cease their usual full-time employment in order to render full-time ADF service).

In the supermarket case study outlined above, access to latent capacity (especially of existing staff working longer/more shifts) was achievable in conditions of temporary surge. A similar approach will generally not be suitable, however, for a sustained response. It may nevertheless 'buy time' to mobilise additional resources via second-order scaling if a sustained response is required. Anecdotally, significant staff turnover rates, including for reasons of 'COVID burnout', remain a challenging legacy of workforce surge during the pandemic.

A final point on capacity: an organisation already at close to full utilisation will have limited scope to access latent capacity (e.g. see the example of health system and ambulance 'ramping' at Annex D). Such an organisation will thus have limited initial capacity to respond to a crisis or shock, which is problematic for organisations with responsibilities to government and obligations to the public to respond, including to shocks. From a leadership perspective, this circumstance carries significant risk.

Insight 3: Business intelligence is a critical enabler of scaling and requires both data and sense-makers

The emergent strategy apparent in the retail supermarket industry's response to the pandemic involved (on the supply side) both redirecting staff effort, and full(-er, -est) utilisation of existing (latent) capacity. The internal ability to apply these two methods, and especially the second, relies on 'knowing the business'—that is, business intelligence with an awareness of the consequences of redirection, and of where latent capacity may lie. This requirement underlines the criticality of *sense-makers*—individuals with a meta-level awareness of their organisation's operations and resources. The case study indicates that both supermarket chains used a combination of sense-makers and data to achieve the requisite levels of business intelligence. Meaningful interpretation of the data was crucial to decision-making on the prioritisation and allocation of resources and effort. For example, digital customer loyalty programs were quickly adapted with an option for selecting vulnerable person status.

Insight 4: Skill sets of existing workforce (generalists versus specialists) and extent of onboarding requirements for new staff can limit scalability

In the main, the supermarkets in the case study generally redirected staff effort *within* existing employment segments. For example, in-store staff were redirected from individual check-out operations to shelf stacking; and warehousing staff were redirected to picking bulk cartons for many stores. This approach was preferred to selective picking for a smaller number of stores. The latter was the favoured pre-COVID practice, because it allows more precise optimisation of orders, deliveries and warehouse space, achieving efficiencies from a business perspective. A more general redirection of staff *between* the in-store and warehouse segments appears to be the exception (to redirection *within* existing employment segments), and presupposes a high degree of generalist skill-set transferability. With role segmentation and specialisation, achieving transferability may require additional training. From a scalability perspective, the greater the transferability of workforce skills, the greater the capacity for redirection. If redirection is not sought (e.g. because both original functions, in-store and warehousing, need to be maintained), then workforce augmentation (e.g. onboarding and training of net new staff) may be required to achieve a scaling effect.

This insight provides a deeper distinction between first- and second-order scalability:

- **First-order scalability** involves redirection of resources and effort within existing segments and is limited by the transferability of the skill sets.
- **Second-order scalability** implies augmentation with additional resources (e.g. new staff), or new/different skill sets.

The scaling implication is that this augmentation will take more time to achieve, as it involves onboarding and training.

Insight 5: Redefining the value proposition (internal performance versus external effects) and challenging paradigms are key levers during a scaling response

The supermarkets' emergent strategy involved, on the demand side, communication with stakeholders to adjust expectations for the standard of service that could be delivered. Relatedly it also involved redefining the organisations' value propositions. In short, managing demand for the delivery of external effects relieved pressure on the internal performance (supply side). A critical insight from this observation is that, during the pandemic, the ability of the retail supermarket industry to (at least partially) manage demand kept the supply-side response within first-order scalability bounds.

In the Defence context, this insight requires careful interpretation. While communication with government, and stakeholder expectation management, are important in ensuring realistic ADF tasking, combat operations may offer limited scope to 'reduce discretionary[76] demand' on ADF services. These scenarios may accelerate leadership decision-making towards a second-order scalability response.

In addition to iteratively firming up the demand signal, the supermarkets' emergent strategy on the supply side involved challenging paradigms of what would have been considered, pre-pandemic, inefficiencies. Examples of these include:

- staff on payroll with a low take-up of shifts
- bulk rather than customised delivery from warehouses to stores
- contingent contracts for access to warehouse space at additional, decentralised distribution nodes.

BAU inefficiency became a necessary scaling redundancy during the pandemic.

Application—the Defence Pandemic Experience

Of course, Defence also initiated a scaling response to the COVID-19 pandemic, albeit in a support role. Prima facie, Operation COVID-19 ASSIST shares some characteristics with the insights provided above, with the exception of Insight 4.

The ADF response was first order (Insight 1), involving redirecting workforce capacity (i.e. redirection of full-time ADF members from training and exercises) and fuller (though still voluntary)[77] utilisation of latent workforce capacity (i.e. call-out of the ADF's part-time Reserve workforce) towards Operation COVID-19 ASSIST (Insight 2). New business intelligence was required to centrally track COVID-19 infection among ADF personnel, including their isolation as necessary, and to maintain visibility of their vaccination status at an individual level (Insight 3). COVID-19 infection control measures (e.g. social distancing and density limits) significantly changed staffing ratios in ways considered inefficient during BAU (Insight 5). However, on Insight 4, in contrast to the supermarkets, the ADF exploited the *high* fungibility in skill set of its existing workforce, with combat soldiers redirected towards aged care—a demonstration of the versatility of the military skill set, and its ability to deliver JIT mission-specific training to meet an emergent contingency. This property of workforce scalability is revisited in Chapter 6, 'Scaling Principles and Their Application to the ADF'. Because further significant augmentation was not required, the ADF could confine its response within first-order scalability bounds.

The fact that the insights drawn from the case study can be applied to the ADF's response to the COVID-19 pandemic indicates that they may offer a robust interpretative tool for scalability, as well as being a basis for further extension.

Summary

For the retail supermarket industry in Australia, the COVID-19 pandemic is a case study in crisis response. The supermarket industry did not plan for a 'shock' to the system, impacting both demand and supply sides, of the scale experienced. It was nevertheless able to develop an emergent strategy in response to this unexpected event. The strategy was temporary. On the one hand, it involved managing demand. Simultaneously, on the supply side, it required the supermarkets to redirect and utilise more fully their existing (latent) capacity and resources, prioritised towards 'essential' aspects of the business. Significant capacity was generated through:

- redirecting the considerable effort dedicated, pre-pandemic, to servicing discretionary needs. This effort was sharpened and focused on utilitarian 'essential' goods and services
- adjusting procedures (e.g. shift and delivery hours).

Collectively, the supermarket response to the pandemic is best understood as an example of first-order scalability—temporary, mostly involving existing capacity, and without significant augmentation in most instances. While the intent of this scaling response—maintaining business continuity and supply of essential goods to the Australian population—was successfully achieved, there were several short- and longer-term consequences, including staff turnover.

It is noted that this case study excludes consideration of cost and profit. While the supermarket industry saw the achievement of an adequate COVID-19 response as essential to its social licence to operate within Australia, there were definitely costs incurred with the new tolerance of 'inefficiencies because they want a guarantee of service'.

Together, Chapters 1 to 5 of this paper have assembled a body of scalability theory and practice. Chapter 6 extends and applies these scalability foundations both in the general sense and for Defence specifically.

Chapter 5: Quick Take-Outs

The case study generated the following five scalability insights from industry with relevance to Defence:

Insight 1: Crisis (within resilience range) will generally initially involve first-order scalability. The barrier to initiating a second-order scalability response is higher in the absence of prior planning. Escalation to second-order scalability is based on crisis:

- magnitude
- duration.

Insight 2: Understanding your organisation's redirectable and latent components of capacity is key to a first-order scaling response.

Insight 3: Business intelligence is a critical enabler of scaling and requires both data and sense-makers.

Insight 4: Skill sets of existing workforce (generalists versus specialists) and extent of onboarding requirements for new staff can limit scalability.

Insight 5: Redefining the value proposition (internal performance versus external effects) and challenging paradigms are key levers during a scaling response.

Australian Army soldiers from 3rd Battalion, The Royal Australian Regiment fire a Javelin Guided Missile during the Direct Fire Weapons Support Course in Townsville Field Training Area, Queensland (Source: Defence image gallery).

Chapter 6: Scaling Principles and their Application to the ADF

The ADF's new capstone concept, Concept APEX, stresses that the ADF is operating in a geostrategic environment of *continuous competition*. Therefore, application of concepts and lexicon from the field of business (where competition is fundamental) is helpful in deriving scaling principles with military relevance. To that end, this chapter extends the conceptual development of scalability to propose scaling principles and metrics directly applicable to the ADF.

This chapter is structured into three sections as follows. Based on the theoretical and practitioner insights developed so far, the first section (6.1) proposes five scaling principles, as an aid to scalability design, response and strategy development. This section extends and applies scalability in a general sense, applicable to all organisations. The remaining two sections extend and apply scalability to Defence. Section 6.2 proposes metrics for the ADF to use to benchmark and monitor its scalability, based on *readiness*, *preparedness* and *procedural potential*. Section 6.3 discusses scalability in the context of Australia's military strategy. The scalability implications of the ADF's new capstone concept, Concept APEX, are considered, alongside a broad assessment of ADF scalability, before areas for potential future work in scalability are outlined.

Scaling Principles

Envisaging both scaling design and scaling response as a journey in strategy development, five scaling principles are presented below, in the order in which they are encountered as an organisation progresses through that journey.

Principle 1: Redirect existing capacity from non-essential to essential, dependent on the scaling imperative

Under resource neutrality, a first response to a scaling imperative requires redirection of currently utilised resources and capacity, based on a reprioritisation. Beyond redirection, scaling requires additional resource allocation.

Reprioritisation applies in both directions—either up- or down-scaling a specific product, output or capability. Reprioritisation may involve physical resources (e.g. staff) or procedural reprioritisation of business rules (e.g. risk tolerance thresholds and delegations).

Where workforce redirection is required, commentating stakeholders can be expected to apply pressure to demonstrate that 'back office' staff are being directed towards the 'front line'. The more complex the scaling event, the more coordination and liaison (i.e. with 'back office' staff) is likely to be required.[78]

Where redirection is required, leaders must make judgments concerning essential versus non-essential activities currently undertaken within the organisation, relative to the scaling imperative. In reality, this judgment rarely involves binary considerations. This is because the critical consequences of redirection are complex and may include cannibalisation of some existing outputs/activities/services. Decision-makers must recognise that any such contingency response should be short term. It generally cannot be sustained for extended periods of time without risk of workforce attrition, capability degradation, and equipment and infrastructure overuse.

Principle 2: Harness latent capacity—business-as-usual inefficiency is necessary redundancy for scaling

Following redirection, scaling next involves utilisation of latent capacity within an organisation—if there is any. By deduction, an organisation or system operating at capacity has limited scope to scale through latent capacity (first-order scalability) and needs to consider second-order scalability to increase or improve performance. As there are costs associated with maintaining latent capacity, the existence of latent capacity within an enterprise can be interpreted as inefficient[79] from a short-term, purely economic rationalist perspective. However, latent capacity represents necessary redundancy if an organisation values the capacity to scale responsively. Paying for latent capacity is an insurance premium against a scaling imperative. It is likely that latent capacity will require leadership sponsorship as a deliberate risk-mitigation measure.

A **scalability mindset** is leadership preparedness to defend latent capacity on the basis of scalability design.

Latent capacity is described below in workforce, equipment and basing terms.

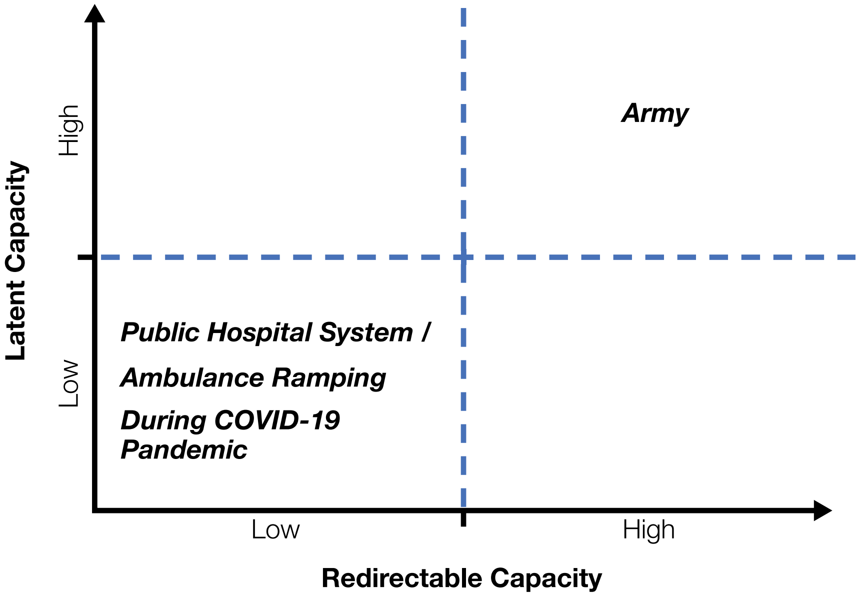
- **Workforce**: There are insourced and outsourced forms of latent capacity. In workforce terms, an example of insourced latent capacity is the ADF's contingent (part-time) workforce, the Reserve. The ADF's personnel capability can be scaled by mobilising this trained workforce for full-time service if required. By contrast, an example of outsourced latent capacity is a contingent option contract for access to a security workforce from a private firm.
- **Equipment**: An example of latent capacity that exists within the ADF is equipment 'fitted for, but not with' certain capabilities. These additional capabilities can be fitted later, should they be required as part of a scaling response. Building in this form of equipment latent capacity is an example of *scalability by design.*
- **Basing**: Chapter 4, 'Scoping and Framing', included the variable 'where' as a key scalability parameter, and discussed how jurisdictional nodes can be activated as a concurrent scaling response. Conceptually, maintaining a network of nodes (even at a rudimentary or 'bare base' level) is a potential source of latent capacity for scaling.

While latent capacity comes with an ongoing cost, the above insource/outsource discussion highlights options to achieve a least-cost latent capacity.

Together, Principles 1 and 2 encompass first-order scalability—that is, redirecting internal resources or utilising latent capacity. Copland describes this process as an organisation 'using its unemployed resources, its idle resources and its resources normally devoted to ordinary investment'.[80] As noted in the industry case study, there are magnitude and duration limits to first-order scalability. If the scaling imperative is a temporary shock to the system, and is within its resilience range, the system can make a temporary surge response using Principles 1 and 2. If the scaling event is sustained, or beyond the enterprise's resilience range in magnitude, second-order scalability may be required, and this involves acquisition and induction of new human, resource or technology capital. Second-order scalability generally requires capital-intensive investment and time, and may build new, permanent capability within the organisation.

The relationship between Principle 1 (redirect existing capacity) and Principle 2 (harness latent capacity) is illustrated in Figure 12.

Figure 12. The capacity relationship between Principle 1 (redirectable capacity) and Principle 2 (latent capacity), plotting the relative positions of the two workforce examples considered: the Australian Army and the public health system (Annex D)



Principle 3: Navigate the scalability–complexity trade-off—simpler capabilities are easier to scale

As explained in Chapter 3, the more complex the value creation process (VCP), the more challenging the scalability task. There are three related aspects of this principle:

- First, other things being equal, new capabilities are more challenging to scale than existing capabilities, as the former involve additional IIS considerations.
- Second, contemporary capabilities tend to be highly integrated and rely on other fundamental inputs.

- Third, from a workforce perspective, the more complex the skill set required to produce and operate a capability, the less fungible that workforce will be. This reality limits Principle 1 (redirection) due to lengthy training requirements. General Stanley McChrystal described this dilemma in the following terms: 'Unfortunately, many of the traits that made our teams so good also made it incredibly difficult to scale those traits across our organization'.[81]

For these reasons, business intelligence is a critical enabler of scalability, especially for more complex VCPs. Business intelligence consists of interdependency mapping of the relevant VCPs, data on the workflows through VCPs, and sense-making. Documentation of standard operating procedures (SOPs) is one example of business intelligence that aids scalability.

Principle 3 crystalises the scalability–complexity relationship as a trade-off. This principle has critical capability consequences, explored further in Section 6.3, 'Scalability and Australian Military Strategy'.

Principle 4: Stay in shape! Avoid unintentionally disproportionate scaling

Disproportionality applies to scaling in two senses. First, the overall shape of an organisation may (intentionally) change through a scaling response (anisometric scaling). For example, a particular capability may prove critical in a particular conflict, so this capability is upscaled relative to others, thus changing the organisation's overall shape. Second, a scaling process may be (unintentionally) disproportionate because elements within an organisation vary in their capacity to scale. The organisation's overall scalability will be limited by the element which is least scalable. Understanding and remedying this is assisted by three ideas: the theory of constraints (ToC), catalysts and absorptive capacity.

- The ToC (covered in Chapter 3) considers scalability as a sequence of binding constraints. The scaling challenge is to understand the spacing of the sequence of constraints within a given process. Spacing is considered in the first instance to be time—i.e., once one binding constraint is remediated, how long is it before the next most binding constraint limits performance, outputs or outcomes? It can be deduced that a closely spaced sequence of constraints is likely to indicate a system approaching its first-order scaling limit, where investments in multiple areas are required to realise increased or improved performance.

- Use of *catalysts*[82] is a concept that can aid a VCP for complex capabilities (and thus scalability). Catalysts can be thought of as producing a multiplier effect in a VCP. Accordingly, the addition of a little (or a few) can generate disproportionate increases in outputs. In military terms, catalysts are known as *enablers*. For example, within a typical Army regiment, there is a regimental quartermaster (RQ) (logistics) function. In a regiment with high logistic requirements, increasing workload will quickly expose the RQ function as the constraint causing a bottleneck in logistic responses, impairing the whole regiment's performance. However, upscaling the RQ team can be a catalyst that alleviates the bottleneck.
- Absorptive capacity[83] sensitises a scaling effort to the receiver of the organisation's scaled product, output or service (Figure 7). For example, a given market size may have a finite capacity to absorb a new product; a given community or local agency may have a finite capacity to absorb a surge workforce; producers may have a limited ability to supply raw material to a firm; distributors may have a limited capacity to deliver finished goods to market. The absorptive capacity generally lies *external* to the organisation undergoing a scaling response and may be related to either the input or the output side of a VCP. Reaching absorptive capacity indicates that an organisation is scaling faster than its external environment.

To overcome the risk of unintentional disproportionality in a scaling process, organisations must ruthlessly pursue the binding constraint (ToC), identify and amplify catalysts, and calibrate the scaling response to the absorptive capacity of the external environment.

Principle 5: Exploit scaling as a transformation opportunity—accelerate and embed positive aspects

Internally, a scaling event can identify (and correct) accumulated procedural inefficiencies in systems and administration (BPAs). Where these inefficiencies do not contribute to latent capacity (Principle 2), they impair scalability. Scaling may present opportunities to accelerate business processes already underway (e.g. digital transformation) and to reform.

Externally, there may be scope, within a scaling imperative, to redefine the relationship between the enterprise and its demand signal, or its key external stakeholders (e.g. government).

As discussed throughout this paper, the task of enterprise leadership and decision-making is to 'know the business' and understand changes in the operating environment. Application of the principles to a scaling response involves at least two critical decisions:

- *containing* a response to within first-order bounds (which requires knowledge of the redirectability of current resourcing, and the extent of latent capability)
- *activating* second-order scalability if circumstances warrant.

The next section considers metrics, specifically in the Defence context, to inform these decisions and the upstream considerations related to scalability design.

Metrics

How can an organisation assess its scalability? As discussed in the introduction, scalability is a statement of an organisation's potential, as well as a statement of its performance. Metrics can assist in assessing both potential prior to and performance during a scaling event. Table 3 and this section outline a framework of proposed scalability metrics in the Defence context based on readiness, preparedness and procedural potential.

Table 3. A framework for scalability metrics

Scalability metric type	Time horizon
Readiness	Immediate and short term ('fight tonight')
Preparedness	Medium term
Procedural potential	Longer term

Readiness

Readiness assessments of scalability include authorities, resources and enablers, outlined below.

- **Authorities**. Examples include:
  - legal authority—e.g. call-out provisions under the *Defence* Act 1903[84]
  - the Australian Government Crisis Management Framework (AGCMF).[85]

Critical questions for authorities include:

  - Are the required sources of authority understood?
  - Are the processes for activation of those authorities understood?
  - Have those processes and authorities been rehearsed or exercised?

- **Resources**. These are critical inputs to the organisation's VCP, including physical, human and financial resources. Consideration here extends to assessment of supporting supply chains, specifically in terms of:
  - reliability: how vulnerable is the supply chain to disruption?
  - resilience: how rapidly can the supply chain recover from general or specific disruptions?

    - redundancy: are there well-developed alternative supply chains available, in case the primary one is disrupted?

While supply chains are most often considered from a physical resource perspective, supply of human resources may also be critical to scalability. In a Defence context, human resource scalability questions include:

- What are appropriate eligibility criteria, given contemporary job requirements, for military, public and industrial workforce?
- How lengthy are the recruitment and onboarding processes for new personnel?
- Is service in these respective workforce components voluntary, or compulsory?
    - In the case of the former, simply attracting sufficient workforce from the wider labour market will be challenging.[86] In the case of the latter, challenges may include motivating and supervising a component of unwilling or reluctant labour.
- **Enablers**. These may include:
    - supporting (cf. authorising) legislation, e.g. the *Defence Reserve Service (Protection) Act 2001*,[87] which protects the employment of Reserve members rendering full-time service
    - access provisions, including:
        - specific technologies (e.g. intelligence-sharing agreements, software licences)
        - air basing and overflight (ABO) in a third-party nation, and status of forces agreements (SOFAs)
    - a geographic network of nodes (Principle 2).

Defence's strategic mobilisation includes a work program of 12 'mobilisation factors'.[88] The scalability readiness metrics of authorities, resources and enablers allow logical grouping of the 12 factors of strategic mobilisation, as shown in Table 4.

Table 4. Grouping the strategic mobilisation factors as scalability readiness metrics

Authorities
• Legal frameworks
• Coordination arrangements
• Sustaining community support
Resources
• Capabilities required
• Personnel
• Infrastructure
• Critical materiel
Enablers
• Partnerships
• Defence communications and culture
• Industry
• Others

Preparedness

Scalability preparedness involves both general and specific measures. The general measure is the response to this question:

- Does the organisation understand its value creation process (VCP), at an individual capability or enterprise level?

By implication, if the VCP and its interdependencies are not mapped, this deficiency can rapidly impair scalability. In terms of specific measures, a scaling event involves analysing the (internal and external) processes likely to be involved in that event, and applying the ToC[89] to identify:

- ease of replicability
- the root-cause source which is limiting the capacity of the process (noting that this may be external to the organisation).

The purpose of root-cause analysis is to identify the binding constraint and to prioritise investment in remediation. In reality, as soon as the binding constraint is remediated, another constraint becomes binding—hence the leadership task of scalability is to resolve the sequence of constraints which limit the capacity of a given system/organisation. Annex D contains a case study example of root-cause analysis to identify the binding constraint in an organisational process: ambulance ramping within the health system.

Procedural Potential

Assessing the procedural scalability of an organisation involves consideration of the following questions. To what extent are:

1. core and other processes *codified*, e.g. through articulated SOPs? This reflects a replicability aspect.
2. process outputs *standardised*, relative to outcome expectations or requirements? This reflects a consistency aspect.
3. simple processes automated?
4. workflows through these processes analysed and tracked, and interdependencies mapped?
5. structures modularised?

Procedural potential involves realising administrative efficiencies which are desirable from a scalability perspective. This is not to be confused with the BAU inefficiencies of Principle 2, which are also desirable from a scalability perspective. Nuanced judgment is required to distinguish whether an inefficiency secures latent capacity.

The concept of procedural potential represents measures within an organisation's control which can be implemented *in advance* of a scaling contingency. This metric is valuable because it helps prepare ahead of crisis. Consistent with this value proposition, the ADF's Concept APEX capstone concept includes the following goal:

Streamline structure, process and procedure. The ADF will maximise effectiveness by simplifying and standardising structures, processes and procedures. Needless variability and intra-organisational seams across the ADF introduce friction and complexity that absorb time, energy and resources. In the procedural dimension, the ADF will be the 'same by default, separate by necessity and similar by exception'.[90]

From a single-service perspective, the Army Business Plan directly links capacity to procedural simplification:

Army needs greater capacity and concurrency to grow new teams quickly in cooperation, competition and conflict. Army will increase capacity through our operating system, and our ability to scale and simplify.[91]

Procedural simplification is a prudent organisational investment in scalability, and can be best supported (beyond statements in business plans and corporate plans) with specific metrics, as outlined above, that are regularly reported.

Collectively, these scalability metrics—*readiness, preparedness and procedural potential*—constitute organisational heuristics (i.e. shortcuts that can be taken in a scaling response, by making use of and building upon institutional features that already exist). In some cases, an organisation may already collect and report on the metrics outlined in this framework. In such cases, the scalability metric framework assists in consolidating currently disparate reporting lines into a coherent enterprise view of scalability. By contrast, scaling ab initio, without an initial organisation, is more challenging.

Benchmarking and Reporting

Collectively, a suite of scalability metrics can form the basis for benchmarking and reporting an organisation's scalability. Scalability is ideally assessed relative to a benchmark. Benchmark examples may include:

- the same organisation, at an earlier point in time
- other similar organisations (e.g. in the same sector)
- effects delivery standard required to meet the demand signal in the external operating environment.

Within the Australian Government, the Department of Finance issues guidance for performance reporting, in terms of principles, application and assessment. Based on this, Defence's corporate reporting to government is well developed, with nested group and service business plans, the Defence Corporate Plan[92] and the Defence Annual Report.[93] This document hierarchy specifies result areas against which performance is reported. While future readiness is included as a result area in Defence reports, there are, however, two barriers to the inclusion of wider scalability metrics:

- A property of or process within an organisation which may have no immediate benefit in the performance of the organisation's activities (but significant longer-term benefit, such as scalability design) may attract less sponsorship from government.
- Performance against such activities (e.g. scalability design, which builds in scaling potential) can be difficult to assess unless and until a specific type of triggering event (e.g. a scaling event) occurs.

Adoption of a scalability mindset by Defence leaders is required to overcome the above barriers and exploit the clear scope to expand future readiness reporting to incorporate scalability.

The following section (6.3) places scalability within the context of Australia's military strategy, as expressed in the ADF's capstone concept, Concept APEX. Section 6.3 also presents a broad assessment of ADF scalability and identifies areas for future work in scalability.

Scalability and Australian Military Strategy

> *To win we had to change … it was about the internal architecture and culture of our force—in other words, our approach to management*[94]

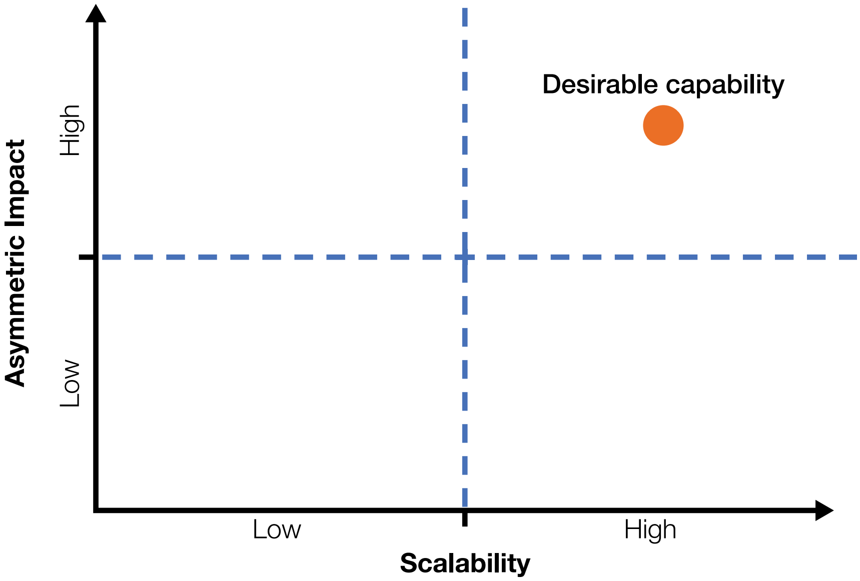
In Chapter 3, the section on the VCP identified positive correlations between simplicity, replicability and ease of scaling. There is, however, an important caveat to the assertion around correlations, and it relates to warfighting utility. Prevailing in warfare, like business, requires a *sustained competitive advantage*.[95] Achieving this advantage requires that the condition of warfare exhibits five collective properties (summarised by the acronym VRINO) of the VCP output:

1. **Valuable**: on a scale from 'useful' to 'indispensable' to the end-user or customer
2. **Rare**: not commonly available
3. **Inimitable**: difficult for competitors to copy
4. **Non-substitutable**: produces a unique effect, not easily achieved via alternative means
5. **Organised**: the VCP is well documented, sufficiently resourced on the input side, and sufficiently supported on the output (distribution/deployment) side.

Military strategy can be broadly defined as the linkage of ends, ways and means to achieve the government's objectives through the application of military power.[96] In an ADF context, the purpose of scalability is not to *scale* (per se), but rather to sustain a competitive advantage long enough to *prevail*—in short, to support Australia's military strategy.

While Scott[97] argues that Australia does not currently have a military strategy, one can be deduced from recent official artefacts, including the ADF's new capstone concept, Concept APEX, and decisions on major capability acquisitions. Concept APEX stresses the supporting concept of asymmetric advantage as the way to prevail over competitors and adversaries. Therefore, VRINO translates into *asymmetric advantage* in a warfighting sense.[98] Applied to the ADF, a given capability may be simple and easy to scale—but unless an asymmetric advantage is generated, it may be of limited utility to the ADF. Thus scalability (in isolation) does not hold primacy for the ADF. This relationship between scalability and asymmetric impact is illustrated in Figure 13.

Figure 13: The relationship between scalability and asymmetric impact. Militarily desirable capabilities are in the top-right of this quad chart.



In recent decades, the ADF has defined its asymmetric advantage primarily in terms of a technology edge—i.e., investment in complex capabilities based on advanced technology. This prioritisation dates from the investment in F-111 under the 1987 Defence White Paper,[99] and continues through to the AUKUS strategic partnership and investment in nuclear-powered submarines,[100] along with most of the complex, high-technology capabilities funded in the Integrated Investment Program (IIP).[101] So, it is evident that the ADF generally seeks to achieve asymmetric advantage via complex, high-technology capabilities. Combining this observation with the insight from the VCP analysis, it is possible to extend Scaling Principle 3 as it applies specifically to Defence.

> *Simpler, easy-to-replicate capabilities are easier to scale and, conversely, complex capabilities are more challenging to scale. Therefore, there is a trade-off between scalability and complexity for ADF capabilities.*

The scalability implications of the ADF's coupling of asymmetry and complexity are significant. The scalability–complexity trade-off implies that the ADF's specific means of achieving asymmetric advantage are difficult to scale. This is problematic in a conflict scenario requiring a capability response that is larger than can be achieved by the ADF's current force in being (FIB). Two relevant examples are:

- **Example 1**: the high quality of the ADF's training is commonly considered an asymmetric advantage. However, if the VCP that generates these trained individuals is complex and lengthy, *training* may be difficult to scale.
- **Example 2**: a complex major platform involves a multi-year manufacturing and assembly process. This *technical capability* is difficult to scale.

Solutions lie in re-examining the concept of asymmetric advantage. There are multiple ways to achieve asymmetric advantage in warfare, and a complex capability consisting of advanced technology is just one way. Similarly, scalability is not necessarily a prerequisite to achieving asymmetric advantage. However, if scalability is ultimately critical to success in conflict, the acme sought are capabilities which are *both* easy to scale *and* offer asymmetric advantage (Figure 13). For example, innovative concepts of employment (CONEMPs) of simpler capabilities may offer asymmetric advantage.

Positioning Scalability within the ADF Concept Hierarchy

Concept APEX refers to 'four actions of applied power: understand, orchestrate, apply effects and sustain'.[102] While scalability is not a warfighting concept ('apply effects') in the first instance, it may be, or become, critical to an *integrated* theory of victory. As such, scalability is likely to be at least partially coupled to warfighting concepts. Three potential applications are offered here:

- A demonstrated preparedness to scale may be an important signal of cost imposition to an adversary.
- Prevailing in conflict is the ultimate VCP, with supporting concepts of asymmetric advantage and scalability as subordinate—offering alternative and complementary ways of achieving this end state.
- Within Concept APEX, the ADF needs to be able to *scale* the four actions of applied power: understand, orchestrate, apply effects and sustain. In this sense, scalability is a unifying umbrella concept in the APEX concept hierarchy.

Differentiating the actions of applied power is significant to scalability in a second respect. When cast as a 'sustain' action, scalability is about force generation—FORGEN. The way a force generates does not have to be the same as the way a force fights. In other words, scalability design and response can be pursued using different methods to an Australian way of war. There are two implications of this insight for Defence. First, this separation opens a wider aperture of solutions to achieve scalability. Second, there is a clear requirement for coherence between decisions on military strategy and force design, and scalability.

Assessing ADF Scalability

Concept APEX's three dimensions of integration—human, procedural and technical—are helpful in assessing the ADF's current scaling constraints and in identifying the factors limiting ADF scalability. These are elaborated below, followed by a worked example:

- The **human** dimension relates to scalability mindset, and specifically to leadership sponsorship of latent capacity as a necessary precondition for the redundancy needed to achieve scalability. The degree of tolerance for latent capacity under BAU is a key indicator of an organisation's scalability mindset.
- The **procedural** dimension relates to the maturity and simplicity of routine business processes within the organisation, and the extent to which they are codified, standardised and automated to minimise internal transaction costs in time and effort.
- The **technical** dimension relates to capability complexity, and specifically dissecting the sources contributing to complexity:
    - Is complexity necessary to achieve the mission, or is it instead an unnecessary BPA[103] warranting procedural reform?
    - Can the same or similar warfighting effects be delivered from simpler suites of capabilities? The latter may have lower unit cost and faster production times.

Applying the above conceptual guidance now to initially assess ADF scalability in practice, Figure 14 revisits the classical example of workforce scalability presented in Figure 1.

Figure 14. Classical workforce scalability revisited—contemporary example

Worked Example

Figure 1 presents the classical example of scalability for workforce. Analysis of workforce from a procedural perspective clearly differentiates first- from second-order scalability, and highlights one factor currently limiting ADF scalability.

First-order scalability:

- Speed and administrative ease with which existing team members can be redirected.

Second-order scalability:

- Speed and administrative ease with which new team members can be inducted.

If the barriers to inducting new/additional capital (human, resources, technology) are high, an organisation may potentially exhibit high first-order scalability but experience significant barriers to second-order scalability. In the ADF context, recent responses to domestic operations (e.g. Operation BUSHFIRE ASSIST 2019–2020, Operation COVID-19 ASSIST, Operation FLOOD ASSIST 2022) have entailed a workforce scaling response involving:

1. redirection of the existing full-time workforce (see Chapter 5)
2. activation of the latent workforce capacity of the Reserve.

Together, these examples demonstrate strong first-order workforce scalability, and high fungibility in workforce skill set. However, lengthy Defence Force Recruiting procedures[104] represent a significant barrier to inducting new ADF members; hence second-order workforce scalability is currently challenging for the ADF.

In contrast to the classical example of soldier scalability achieved in World War I and World War II, contemporary ADF capabilities are challenging to scale. This is due to *complexity*, both in terms of the production process for capabilities involving advanced technology, and in terms of the interdependencies involved in generating and delivering networked effects.

Even considering the single capability of 'soldier', contemporary entry and training standards are high and lengthy relative to the classical example. Scalability is also segmented by capability and by operating system.

This broad initial assessment of ADF scalability across the human, procedural and technical dimensions suggests that the ADF exhibits high first-order scalability but may be challenged to achieve second-order scalability.

Metrics Revisited

Chapter 1 noted that, across all sectors, pilot programs for new capabilities privilege measures of performance and tend to omit measures of scalability. A key recommendation arising from this analysis is that candidate projects for Defence's acquisition and sustainment budget, the IIP, include scalability metrics alongside performance metrics.

Table 5 draws together recommendations from Section 6.2, 'Metrics', and this section, recognising the parallel opportunities to embed scalability in both the corporate enterprise and ADF warfighting facets of the Australian Defence Organisation. Adopting a framework of scalability metrics for existing and proposed capabilities as outlined in Table 5 is the first step in converting qualitative statements on relative scalability into quantitative statements which can assess the adequacy of Defence scalability relative to strategic warning time. Once the metrics are benchmarked, an **enterprise scaling strategy** can outline subsequent steps and measure organisational progress towards enhanced scalability, to the extent that this is considered desirable in support of Australia's military strategy. Several initiatives already underway within the Defence portfolio under the Defence Transformation Strategy[105] are particularly relevant to scalability, especially procedural potential. Scalability forms a nucleus around which clear military intentionality can crystalise, alongside administrative efficiency goals often underpinning recent organisational reform. Scalability offers a unifying construct which can bring together the military and administrative functions of the portfolio.

Table 5. Applying 'performance' and 'scalability' metrics to the military and administrative functions of the Australian Defence Organisation (ADO)

Organisational application	**Performance** Metric	**Scalability** Metric
Enterprise corporate reporting	• Current performance metrics • Current capability and acquisition programs	Opportunities: • Report ADO scalability metrics • How scalable are these capabilities and acquisitions?
ADF warfighting concept hierarchy	E.g. Concept APEX's four actions of applied power: • Understand • Orchestrate • Apply effects • Sustain	Opportunity: How scalable are these four actions?

Future Work

There are opportunities to further develop the theory, practice and institutional reach of scalability within Defence:

1. On the theoretical front, there is scope for further exploration of the links between scalability and two fields of literature: systems theory (resilience) and complexity theory (interdependencies). This may generate deeper insights useful for organisational scalability.
2. On the practice front, case studies of significant contemporary second-order scalability in large organisations would be instructive.
3. On the institutional front, Defence may consider adopting a framework of scalability metrics for existing and proposed capabilities, and developing a scaling strategy to complement military strategy development.

Scaling Strategy

The initial conceptual model presented in Part 1 of this paper structures scalability into first and second orders. First-order scalability is the extent to which an existing organisation (or system) can scale, *within its existing resourcing footprint*, by redirecting resources towards more *essential* functions. Second-order scalability occurs when the existing organisation (or system) has reached its full capacity constraint, and external augmentation is required to scale further. For the ADF, this means recourse to government.

To address the opening premise of this paper, an organisation can enhance its scalability by applying scalability design and response functions to develop a scaling strategy. Part 2 has offered practitioner-focused proposals focused on scalability with specific reference to:

- scoping and framing considerations
- industry insights (through case study)
- principles
- metrics for benchmarking and reporting.

Collectively, these observations can inform development of a scaling strategy. Absent an a priori scaling strategy, an organisation's typical response to a scaling imperative involves predominantly first-order scalability. This effort is operational in nature, redirecting currently utilised capacity and mobilising latent capacity towards new priorities. However, organisations are limited in their first-order scalability, and beyond this, second-order scalability is required. As has been seen, second-order scaling requires the commitment of significant additional capital investment and time to transition the state of an organisation sufficiently to significantly augment its capacity.

While second-order scalability is a design function, there are also design aspects involved in building in latent capacity (first-order scalability) within organisations. Sponsorship by leaders—and specifically the adoption of a scalability mindset—is required to defend scalability design against managerialist instincts, which tend to view latent capacity as inefficient under BAU. There is a tension between scalability design and response, reflecting an organisation's view on the extent to which the future can be predicted. For organisations which are inherently likely to require scaling in the future, developing a scaling strategy is a wise investment.

Implications

Organisational scalability is best understood as a *craft*—combining both science (requiring technical expertise) and the art form of ‘knowing the business’ (understanding both the organisation under study, and the unique context of the environment in which it operates). Achieving scalability for the ADF requires integration of these two types of knowledge at the meta level. This places a premium on the ‘sense-makers’—those individuals (or leadership teams) who can discern and articulate internal organisational dynamics and scaling constraints and identify external supporting efforts and resourcing (if/as required) during a scaling response.

Appreciating first- and second-order scalability enables an organisation to identify the capacity investments required to enhance its scalability:

- **First-order scalability requires internal investment in people and processes.** A ‘scalability mindset’ can predispose teams to rapidly receive and assimilate new members during a scaling phase. Opportunities exist to scale processes in, for example, selecting, onboarding and training new members. Process enablers for enhanced scalability include reconsidering risk thresholds (e.g. ‘How long could/should an initial recruit course be if there is a rapid scaling requirement?’) and automation (i.e. transitioning manual tasks towards digitisation—e.g. online delivery of training to much larger groups).
- **Second-order scalability requires investments in the *interconnectors* between the organisation and its ‘adjacent ecosystem’**. For the ADF, the ‘adjacent ecosystem’ consists of government architecture and the wider national support base (e.g. Defence industry) that can be harnessed to assist the Defence portfolio to rapidly scale. Interconnectors include the leaders and individuals who manage an organisation’s relationships externally. These interconnectors may be required to pitch and win business cases for additional resources from a sponsor to fund an upscaling, for example.

A critical deduction follows from the above structured understanding of scalability. Specifically, an organisation’s capacity to achieve scalability is premised on its capacity to recognise whether the scalability effect required is first- or second-order, as these invoke different resolution pathways. The threshold between the two requires acute attention to existing capacity constraints—again, contextualised to a specific organisation.

Sense-Making and Influence

Applied in an ADF context, officers of O6 (Colonel-equivalent) rank represent some of the ADF's sense-makers, as these individuals:

a. generally have acquired posting experience across different parts of the Defence organisation

b. have sufficient rank within senior leadership teams to articulate scalability issues and exert influence towards their resolution

c. are closely enough involved in the minutiae of Defence's operations, actions and activities (OAA) to develop a sense of 'where the business is at' and, importantly, where the constraints lie.

Applied more broadly, organisational sense-makers can be found among the often-maligned 'middle management' of organisations. To what extent can overall organisational sense-making be codified? Is it possible to capture the corporate knowledge held by an organisation's scalability sense-makers into an information management system, for example? The view that currently prevails within the business literature is that, beyond a certain level, codifying expert knowledge into ICT or automated 'expert systems' is prohibitively costly in terms of translation effort, and of limited value due to the dynamic nature of this knowledge.[106]

In the context of ADF's 'adjacent ecosystem', much of the requisite expert knowledge is relationship based, rather than procedural. Where relationships are influential, unlocking a scalability constraint may involve identifying the point of influence on a scalability issue (i.e. the critical human decision-maker) and finding (often lateral) means of advocacy to nudge[107] the decision-making towards actions, decisions and resource allocation that may help unlock an otherwise binding constraint on an organisation attempting to scale. A critical difference between the examples provided in business literature and the circumstances of Defence is that the former often assumes that the majority of the critical organisational decisions can be made by the organisation's own leadership team. This is not always true for the Defence portfolio, where the point of influence may lie outside the organisation—for example, partnered forces, other government agency stakeholders, or politicians.

A key deduction from this analysis is that enhanced ADF scalability involves two initial steps:

1. Accurate, contextual sense-making, to identify the organisation's scalability constraint
2. Effective advocacy at the point of influence, to achieve appropriate resourcing decision-making. This is bounded by the ethics of Moore's theory of public value (Figure 9).

The third step is implementing the result of positive decision-making: rapidly converting additional resourcing into scaled delivery effects. This topic is the subject of Part 2 of this paper.

Chapter 6: Quick Take-Outs

Table 6. Summary map of industry insights and general principles generated in this paper, with arrows indicating the journey in developing a scaling strategy

Scalability	Insights	Principles
First-order scalability	**Insight 1—crisis response**: Crisis (within resilience range) will generally initially involve first-order scalability. The barrier to initiating a second-order scalability response is higher in the absence of prior planning.	**Principle 1**: Redirect existing capacity from non-essential to essential, dependent on the scaling imperative.
First-order scalability	**Insight 2—capacity**: Understanding your organisation's redirectable and latent components of capacity is key to a first-order scaling response.	**Principle 2**: Harness latent capacity—business-as-usual inefficiency is necessary redundancy for scaling.
First- and second-order scalability interface	**Insight 3—business intelligence**: Business intelligence is a critical enabler of scaling and requires both data and sense-makers.	**Principle 3**: Navigate the scalability–complexity trade-off—simpler capabilities are easier to scale.
First- and second-order scalability interface	**Insight 4—workforce fungibility versus segmentation**: Skill sets of existing workforce (generalists versus specialists) and extent of onboarding requirements for new staff can limit scalability.	**Principle 4**: Stay in shape! Avoid unintentionally disproportionate scaling. Use the theory of constraints, catalysts, and absorptive capacity to maintain scaling momentum.
Scaling as transformation	**Insight 5—internal performance versus external effects**: Redefining the value proposition and challenging paradigms are key levers during a scaling response.	**Principle 5**: Exploit scaling as a transformation opportunity—accelerate and embed positive aspects.

Outgoing Commanding Officer His Majesty's Australian Ship Canberra, Captain Jace Hutchison, Royal Australian Navy speaks to the ship's company during a change of command ceremony while alongside Fleet Base East, NSW (Source: Defence image gallery).

So What for Defence? The Scalability Top 10

This Scalability Top 10 is not exclusively applicable to Defence. How many in the list below also apply to your organisation?

1. **Develop a scaling strategy**, which includes both design (planning) and response (operational) aspects. The 2023 Defence Strategic Review[108] recommends that Defence undertake 'Accelerated Preparedness in Competition'. Accelerated preparedness equates to scalability, and a scaling strategy can deliver accelerated preparedness as an outcome.

   An optimal scaling outcome is based on both design (prior to) and response (during) a scaling event. Without a scaling strategy, Defence loses the asymmetric advantage of design, and is forced straight into 'response mode' during a scaling event. Scalability design helps build in latent capacity (first-order scalability) and pre-positions the organisation for second-order scalability if required.

2. **Include explicit scalability metrics in corporate reporting** (at appropriate classification levels where required). There is clear scope to expand future readiness and preparedness reporting to incorporate scalability. This signals value to stakeholders.

3. **Assess potential new capabilities** both for asymmetric impact *and* for scalability.

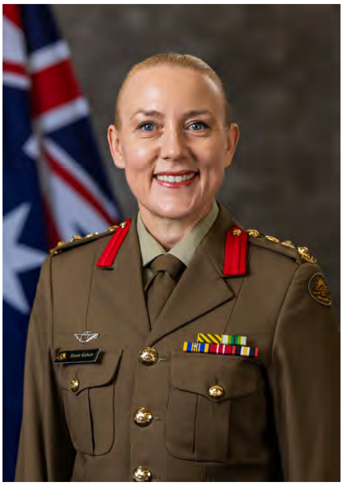
4. **Cultivate a scalability mindset**. The role of scalability leaders is to:
    - in competition—advise on and advocate for scalability design features
    - in crisis and conflict—ruthlessly 'find and fix' the sequence of binding constraints that will otherwise stall a scaling response.
5. **Invest in internal business intelligence**. How well do we understand our internal processes? Our latent capacity?
6. **Cultivate organisational sense-making**. This requires both data (the business intelligence generated above) and skilled humans to meaningfully interpret it. While artificial intelligence is part of the solution, it is not the complete solution.
7. **Invest in internal business process efficiencies** that promote productivity (e.g. automation of routine administration)—but retain the inefficiencies which constitute latent capacity for future scaling (e.g. the training overhead of part-time workforce components). Ideally, efficiencies gained in business improvement are reinvested into building latent capacity. In workforce terms, the Reserve is the ADF's critical latent capacity, and efforts to build the Reserve are an investment in future workforce scalability.
8. **Invest in critical enablers early**. These will be among the first binding constraints in a scaling response. A comprehensive analysis of fundamental inputs to capability (FIC) can identify the sequence of binding constraints. Defence is challenged to address these binding constraints because (among other things) at an organisational level the single services don't 'own' all their FIC.
9. **Invest in external relationships** with partner portfolios, with government, with Defence industry domestically, and with like-minded partner nations internationally. This adjacent ecosystem constitutes the absorptive capacity that Defence may require to scale. For Defence, the scaling constraints often lie outside the organisation itself. Constraints may include the government's budget allocation to Defence, caps on personnel numbers, or an irreducible manufacturing time in Defence industry.

10. **Consider scalability from both the force provider and the force employer viewpoint**s. Scalability is not just about the inputs to the value creation process (VCP)—the force provider view. Scalability is ultimately about scaling *delivery effects*—the force employer view. There are multiple ways to scale delivery effects:

    - People
    - Technology
    - Regional relationships.

    Incorporating scalability into the ADF's warfighting concepts will ensure exploration of all possible ways and means of prevailing in future conflict.

About the Author



Colonel Renée Kidson CSM is currently Colonel Domestic Operations and Plans in Headquarters 5th Brigade of the Australian Army. Renée is a dual-career exemplar. Her military career has spanned 27 years to date, which she has proudly served as a Reservist concurrently with her civilian executive career. In her ADF capacity, she has served on multiple deployments internationally and domestically, most notably as Task Group Commander in New South Wales during Operation BUSHFIRE ASSIST 2019–2020, for which she received the Conspicuous Service Medal. Other career highlights include serving as Commanding Officer of the 5th Engineer Regiment, and as the founding Director of the Army Research Centre.

A scientist and economist by training, Renée's current civilian role is Executive Director Strategy in Navy Fleet Headquarters in Sydney.

Renée has six degrees, including a PhD in Science from Trinity College Cambridge, and a Master of Economics (Honours) from the University of Sydney. In 2022 she completed the MBA at Deakin University as part of the Defence and Strategic Studies Course. For her MBA work, including her thesis on Defence scalability, Renée was awarded the Dean's Merit List academic prize.

From 2021–22, Renée was the Australian Army's Director Scalability, and in that role founded the theory and practice components of scalability presented in this work.

This is Renée's third publication; her previous works are *Force Design in the 1990s: Lessons for Contemporary Change Management* (2017) *and Scaling the Force* (2022).

When not engaged in her dual careers, Renée enjoys entertaining with old-fashioned home cooking, traditional wool yarncraft, and spending time in the great Australian outdoors and at her modest beach house.

The author is definitely not scalable; however, the insights presented in this paper definitely are. Renée extends an open invitation now for others to contribute to and expand this field of organisational knowledge.

Annex A: Literature Definitions of Scalability

	Definition	Discipline		Source
1	The ability of something, especially a computer system, to adapt to increased demands.	General	General	‘Scalability’, dictionary.com, https://www.dictionary.com/browse/scalability, accessed 19 June 2022
2	A scalable system has increasing performance with increasing system size … preferably super-linear increases in system performance with scale.	ICT (computing and robotics)	Quantitative	H Hamann and A Reina (2022), ‘Scalability in computing and robotics’, *IEEE Transactions on Computers*, 71(6), 1453–1465, https://doi.org/10.1109/TC.2021.3089044, p. 1453
3	The ability to function well as system size or needs increase, and the ability to take advantage of increases in system size. This is important, as it is common for scaled systems to offer diminishing returns on added assets after they reach a certain size. In multi-UAV109 systems, a concrete definition is that scalability means that ‘group numerality does not result in a drastic change in the performance of the group’.	ICT (robotics)	Quantitative	J Humann and KA Pollard (2019), ‘Human factors in the scalability of multirobot operation: a review and simulation’, *2019 IEEE International Conference on Systems, Man and Cybernetics* (SMC), 6–9 October 2019, p. 700
4	The potential of a system to handle a certain amount of work and its ability to grow seamlessly as needed.	ICT (internet)	Quantitative	C Dhall (2018), *Scalability Patterns: Best Practices for Designing High Volume Websites* (Apress), 1. Introduction, ‘Concepts’

	Definition	Discipline		Source
5	The capacity of programs and interventions to increase in reach and impact.	Social sciences (public policy)	Qualitative	W Hsieh, R Wickes and N Faulkner (2022), ‘What matters for the scalability of prejudice reduction programs and interventions? A Delphi study’, *BMC Psychology*, 10(1), 1–14, Results, ‘Definition of Scalability’
6	The capacity of an individual intervention to be scaled up.	Social sciences (health)	Qualitative	S Calnan, K Lee and S McHugh (2022), ‘Assessing the scalability of an integrated falls prevention service for community-dwelling older people: a mixed methods study’, *BMC Geriatrics*, 22(1), 17, Calnan et al. (2022), p. 1
7	Scalability of multi-agent systems (MAS) refers to the ability of the MAS to gracefully change performance under variation of different parameters. On the one hand, we can distinguish quantitative scalability which depends on quantitative changes in parameters like resources and number of agents. Qualitative scalability depends, on the other hand, on scaling the complexity of social relationships from simple interactions to creating organisations or even further to forming artificial societies with increasing agent complexity, i.e. improving the abilities of agents to deal with complex situations, as well as increasing problem complexity. While qualitative scaling is concerned with increasing (social) complexity requiring new dimensions in perception and decision making, quantitative scalability tackles the problem how goals can be achieved under the constraints imposed by a growing population.	Social sciences (socionics)	Qualitative	K Fischer and M Florian (2005), ‘Contribution of socionics to the scalability of complex social systems: introduction’, in K Fischer, M Florian and T Malsch (eds), Socionics: Lecture Notes in Computer Science, vol. 3413 (Springer), https://doi.org/10.1007/11594116_1, pp. 7–10

	Definition	Discipline		Source
8	The ability of a network to expand such that it operates with acceptable effectiveness.	Social sciences (business)	Qualitative	Jablonski (2017),). *Sustainability and Scalability of Business: Theory and Practice*. Nova Science Publishers, Inc. p. 13
9	Dissemination of change across different contexts … Scalability may mean an innovation staying at one level, such as transferring a second-order technological innovation from school to school or scaling from a second-order school change to a third-order district-level initiative.	Social sciences (education)	Qualitative	SK Howard, L Schrum, J Voogt and H Sligte (2021), 'Designing research to inform sustainability and scalability of digital technology innovations', *Educational Technology Research and Development*, 69(4), 2309–2329, https://doi.org/10.1007%2Fs11423-020-09913-y, p. 2311

Annex B: Second-Order Scalability—an Urban Road Network Example

The system performance of a fixed road network is a physical infrastructural example, where an increasing number of vehicles results in initially diminishing system performance (route travel times), and finally declining system performance, where road congestion leads to lengthy traffic jams and gridlock. Typical scaling solutions (scaled in ascending order of sophistication) to this repeated urban dilemma include:

1. **Augment the existing road network**, e.g. widen the carriageway for additional lanes, duplicate the existing carriageway, construct additional underpasses and overpasses.
2. Develop and encourage use of **alternative transport systems** such as bus and rail networks.
3. **Manage demand**, through incentivising commuters to shift trip timing towards the shoulders rather than the peak periods.
4. **Encourage work from home**, especially during major events or in response to unplanned contingencies (e.g. a natural disaster affecting the road network, a pandemic).
5. **Redesign the need to travel** for a given metropolis, e.g. encouraging development of 'satellite cities' within a region, reducing commuter numbers, commuting times and distances.

Annex C: ADF Scaling Analogy—C2 Systems

Prima facie, the ADF is well positioned to achieve horizontal scalability, understood as adding (or subtracting) ‘units’. Two scales are considered here: task organisation for a limited-scope operation, involving part of the ADF; and a perspective from the ADF as a whole.

Task-organised operation Typical military command and control (C2) structures are conducive to expanding or contracting horizontally, where a headquarters node can accommodate additional ‘units’(e.g. multiple task units reporting to a task group headquarters) up to a certain point. Beyond this point, leading and managing the structure may become unwieldy, either because the ‘span of command’ is too wide, or because the functional or geographic range of units becomes too diverse. The symptom (or metric) indicating that this point has been reached is headquarters response times: headquarters staff overloading slows down routine processes. This analogy reflects the concave curve of first-order scalability shown in Figure 2.

Once this point is reached, a vertical scaling solution is required: either upscaling the headquarters size (e.g. from a task group to a task force headquarters); or the structure may be divided into several task groups. This analogy reflects a physical type of state transition (second-order scalability), shown in Figure 2. A different type of state transition may occur if (for example) an AI-enabled targeting system is implemented within the headquarters, which automates a previously manual process. This technical enabler can also speed up headquarters responsiveness.

Total ADF If the ADF as a whole is considered, there are a limited number of ‘units’ within the existing force: the force in being (FIB). Beyond this point, additional units must be recruited and trained, and the limiting factor (constraint) for scalability is the speed with which these additional units can be generated.

Therefore, the highly structured C2 systems of the military confer strong theoretical scalability, until the size constraint of the FIB is reached. Beyond this point, other enablers (resourcing, recruiting and training systems) become the scaling constraint.

Annex D: Case Study—Root-Cause Analysis

The **purpose** of root-cause analysis is to identify the binding constraint in a process. The critical **importance** of root-cause analysis is because the ToC holds that at any given time, only one constraint acts as binding on the performance of a specific process. Further, the ToC holds that investing effort in anything other than the binding constraint is futile in the sense that it will neither alleviate the binding constraint nor improve performance. Therefore, root-cause analysis is useful in prioritising investment.

Worked Example: Health System and Ambulance 'Ramping'

During the COVID-19 pandemic, the Australian population became familiar with a new phenomenon: ambulance 'ramping'.[110] A basic root-cause analysis of this process is below.

Observed phenomenon: some reported deaths from failure of the ambulance service to respond in time to emergency calls from/on behalf of acutely unwell people.

Why?

- For the sample of observed cases:
  - Were the emergency calls answered?
    - No: this suggests a capacity constraint may exist in the emergency service call centre
    - Yes: this suggests a capacity constraint may lie elsewhere
  - Were there ambulances available?
    - No
  - Why were the ambulances not available?
    - Ambulances containing unwell people were queueing at hospitals—known as 'ramping'
  - Why were the ambulances queueing at hospitals?
    - Hospitals were unable to admit patients

- Why were hospitals unable to admit patients?
  - Were there available beds? (Yes)
  - Were there available medical staff (No)

Consequence: In this case, if the binding constraint is medical staff in the hospital system, investment in more emergency service call centre staff, additional ambulances or additional hospital beds will not necessarily alleviate the binding constraint. This case is an example of a complex system, where the reported observation may not indicate the root-cause constraint, and may instead represent a symptom.

Measure of performance (MoP): In this example, ambulance service statistics indicate ambulance response times have significantly increased, including the frequency of instances of waiting times in the 'unacceptable' category for acute illness types.

This MoP (metric) is an indicator of a system approaching capacity constraints (first-order scalability limit). Conceptually, state transition is required to achieve second-order scalability and significantly augment existing capacity.

Endnotes

1 Australian Government (2023), National Defence: Defence Strategic Review (Commonwealth of Australia), https://www.defence.gov.au/about/reviews-inquiries/defence-strategic-review

2 Chapter 3 defines mobilisation. For the immediate purpose here, it is important not to conflate mobilisation with conscription or mandatory military service. In Australia's case, while conscription was debated, it was not in the end required, as sufficient volunteers enlisted. In fact, industrial workers became Australia's initial 'binding constraint'.

3 'First World War 1914–18', Australian War Memorial, https://www.awm.gov.au/articles/atwar/first-world-war, accessed 26 October 2022.

4 'Second World War, 1939–45', Australian War Memorial, https://www.awm.gov.au/articles/second-world-war, accessed 26 October 2022.

5 J Grey (2008), *A Military History of Australia*, 3rd edn (Cambridge University Press).

6 RL Kidson (2017), *Force Design in the 1990s: Lessons for Contemporary Military Change Management* (Australian Army, Commonwealth of Australia).

7 Department of Defence (DoD) (2013), *ADDP 00.2 Preparedness and Mobilisation*, Executive Series (Australian Government), para. 5.2, p. 5-4.

8 SP Huntington (1957), *The Soldier and the State: The Theory and Politics of Civil-Military Relations* (Vintage Books).

9 H Strachan (2020), 'Strategy and democracy', *Survival*, 62(2), 51–82, https://doi.org/10.1080/00396338.2020.1739949. See also B Heuser (2022), War: *A Genealogy of Western Ideas and Practices* (Oxford University Press), https://doi.org/10.1093/oso/9780198796893.001.0001

10 J Coyne, G Savage and M Shoebridge (2021), 'New beginnings: rethinking business and trade in an era of strategic clarity and rolling disruption', 14 September, Australian Strategic Policy Institute (ASPI), https://www.aspi.org.au/report/new-beginnings-rethinking-business-and-trade. See also D Horne (1964), *The Lucky Country: Australia in the Sixties* (Penguin Books).

11 E.g. EM Rogers (1962), *Diffusion of Innovations* (Collier Macmillan Inc.).

12 C Adam, B Gunasingham, J Graham and S Smart (2017), *Introduction to Corporate Finance: Asia-Pacific Edition*, 2nd edn (Cengage).

13 ‘The first, the supreme, the most far-reaching act of judgment that the statesman and commander have to make is to establish the kind of war on which they are embarking’. C von Clausewitz (1976 [1832]), *On War*, ed. and trans. Michael Howard and Peter Paret (Princeton University Press), p. 88.

14 C Isaacson (2014), *Understanding Big Data Scalability*, Big Data Scalability Series, Part I (Pearson).

15 Ibid.

16 ML Abbott and MT Fisher (2011), *Scalability Rules: 50 Principles for Scaling Web Sites* (Addison-Wesley).

17 AK Tyagi (2021), *Data Science and Data Analytics: Opportunities and Challenges* (CRC Press LLC).

18 TV Guy, M Kárný and DH Wolpert (eds) (2015), *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability* (Cham: Springer).

19 W Hsieh, R Wickes and N Faulkner (2022), ‘What matters for the scalability of prejudice reduction programs and interventions? A Delphi study’, *BMC Psychology*, 10(1), 1–14.

20 C Dhall (2018), *Scalability Patterns: Best Practices for Designing High Volume Websites* (Apress), Part 1: Introduction.

21 H Liu (2009), *Software Performance and Scalability: A Quantitative Approach* (Wiley-Blackwell), p. 1.

22 CHT Arteaga, A Ordoñez and OMC Rendon (2020), ‘Scalability and performance analysis in 5G core network slicing’, *IEEE Access*, 8(142), 86–100, https://doi.org/10.1109/ACCESS.2020.3013597; Hsieh et al. 2022.

23 J Humann and KA Pollard (2019), ‘Human factors in the scalability of multirobot operation: a review and simulation’, *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 6–9 October 2019, p. 700.

24 Dhall 2018.

25 E.g. Isaacson (2014) refers to scalability as ‘elastic’.

26 H Hamann and A Reina (2022), ‘Scalability in computing and robotics’, *IEEE Transactions on Computers*, 71(6), 1453–1465, https://doi.org/10.1109/TC.2021.3089044, p. 1456; Humann and Pollard 2019.

27 Arteaga et al. 2020.

28 P Roquero and J Aracil (2021), ‘On performance and scalability of cost-effective SNMP managers for large-scale polling’, *IEEE Access*, 9, 7374–7383, https://doi.org/10.1109/ACCESS.2021.3049310

29 Hamann and Reina 2022.

30 Ibid.

31 Humann and Pollard 2019.

32 Ibid., p. 700.

33 Liu 2009, p. 2.

34 Abbott and Fisher 2011.

35 ‘When done right, scaling out is a long-term solution for almost any application [ed: cf. database] performance issue; in other words, it’s a real and permanent fix.’ Isaacson 2014, Chapter 5: ‘Scaling your application’, ‘Scaling out’.

36 J Doran (2020), '*How Phorest managed scalability challenges with a continuous improvement mindset*' (O'Reilly Media, Inc.), https://learning.oreilly.com/videos/how-phorest-managed/0636920460053/0636920460053-video330004

37 E.g. by Humann and Pollard (2019) and Hamann and Reina (2022).

38 Dhall 2018.

39 J Ponge and M Little (2019), 'Scalability and resilience in practice: current trends and opportunities', *38th Symposium on Reliable Distributed Systems (SRDS)*, 1–4 October 2019.

40 AI Fajri and F Mahananto (2022), 'Hybrid lightning protocol: an approach for blockchain scalability issue', *Procedia Computer Science*, 197, 437–444, https://doi.org/10.1016/j.procs.2021.12.159

41 ME Latoschik, F Kern, JP Stauffert, A Bartl, M Botsch and JL Lugrin (2019), 'Not alone here?! Scalability and user experience of embodied ambient crowds in distributed social virtual reality', *IEEE Transactions on Visualization and Computer Graphics*, 25(5), 2134–2144, https://doi.org/10.1109/TVCG.2019.2899250

42 Q Zhou, H Huang, Z Zheng and J Bian (2020), 'Solutions to scalability of blockchain: a survey', *IEEE Access*, 8, 16440–16455, https://doi.org/10.1109/ACCESS.2020.2967218

43 E.g. T Dunning and E Friedman (2021), *AI and Analytics at Scale: Lessons from Real-World Production Systems* (O'Reilly).

44 Arteaga et al. 2020.

45 J Petrović and P Pale (2021), 'Achieving scalability and interactivity in a communication skills course for undergraduate engineering students', *IEEE Transactions on Education*, 64(4), 413–422, https://doi.org/10.1109/TE.2021.3067098, p. 414.

46 T Manderson, M Weaver, M Sinsky, M Cook, G Homsy, E Nolasco, R Legge, A Srivatsan and A Eustace (2021), 'The convergence of efficiency and scalability in ocean data platforms', *OCEANS 2021*, San Diego, 20–23 September 2021.

47 A Gulati, K Ganguly and H Wardhan (eds) (2022), *Agricultural Value Chains in India: Ensuring Competitiveness, Inclusiveness, Sustainability, Scalability, and Improved Finance* (Springer).

48 O Al-Ubaydli, JA List and D Suskind (2020), '2017 Klein Lecture: The science of using science: toward an understanding of the threats to scalability', *International Economic Review*, 61(4), 1387–1409, https://doi.org/10.1111/iere.12476

49 Dhall 2018.

50 Supplee and Kane (2021). The realities of scaling within evidence-based policy. *Behavioural Public Policy*, 5(1):90-102.

51 S Calnan, K Lee and S McHugh (2022), 'Assessing the scalability of an integrated falls prevention service for community-dwelling older people: a mixed methods study', *BMC Geriatrics*, 22(1), 17, https://doi.org/10.1186/s12877-021-02717-6, p. 11.

52 Kidson 2017.

53 In some cases, a scaling event may require an organisation to deliver its existing products and services, but in a changed proportion (e.g. less in-store service, more home delivery). This represents a change in organisation shape—the scaling ratio. See 'Scaling Parameters' section in this chapter.

54 E.g. Isaacson 2014, Chapter 2: 'Why databases slow down'.

55 A Watkins and S May (2021), *Innovation Sucks! Time to Think Differently* (Taylor & Francis Group), p. 151.

56 JM Perloff (2007), *Microeconomics*, 4th edn (Pearson Education), Appendix 6B.

57 After economist Joseph Schumpeter's 'creative destruction': JA Schumpeter (1942), *Capitalism, Socialism and Democracy* (Routledge).

58 M Moore (1995), *Creating Public Value: Strategic Management in Government* (Harvard University Press); M Moore (2021), 'Creating public value: the core idea of strategic management in government', *International Journal of Professional Business Review*, 6, 219, https://doi.org/10.26668/businessreview/2021.v6i1.219

59 Huntington 1957.

60 E Goldratt (1999), *Theory of Constraints* (North River Press). See also HW Dettmer (2007), *The Logical Thinking Process: A Systems Approach to Complex Problem Solving* (American Society for Quality, Quality Press).

61 A technical, practical application of ToC is Amdahl's Law (Liu 2009, p. 97), an empirical scaling heuristic which determines how fast an overall system will speed up if investment is made in speeding up a given subsystem component. The relation is a product function, contingent on the 'impact factor' of that subsystem to the whole.

62 Dr Jeff Chamberlain, Chair of Business Process Improvement at Deakin University, 20 June 2022, pers. comm.

63 DoD 2013, p. 7.

64 DoD (2022), 'Defence strategic mobilisation talking points', 24 August, Department of Defence, Australian Government.

65 P Harmon (2014), *Business Process Change: A Business Process Management Guide for Managers and Process Professionals*, 3rd edn (Morgan Kaufmann Publishers).

66 ADF (2024), *Concept APEX: Integrated Campaigning for Deterrence. The Australian Defence Force's Capstone Concept*, 2nd edn (Department of Defence, Australian Government).

67 In the ADF military context, 'Scoping and Framing' is the first step in the Joint Military Appreciation Process (JMAP).

68 E.g. Kidson 2017.

69 E.g. in the Australian water and energy utilities. See N Hughes, A Hafi and T Goesch (2009), 'Urban water management: optimal price and investment policy under climate variability', *Australian Journal of Agricultural & Resource Economics*, 53(2), 175–192, https://doi.org/10.1111/j.1467-8489.2007.00446.x; and R Mitelman (2012), 'Sustainable energy development: the role of demand-side management in Australian energy policy', *National Economic Review*, 67, 44–52, https://doi.org/10.3316/informit.018542362416785

70 This is the 'Proteus Problem', where militaries wrestle with a constantly shape-shifting adversary, articulated by S McChrystal, T Collins, D Silverman and C Fussell (2015), *Team of Teams: New Rules of Engagement for a Complex World* (Portfolio Penguin).

71 Harmon 2014.

72 See Chapter 3 for a full discussion of ToC.

73 RL Kidson (in press), 'Scaling the force: Reserve mobilisation for domestic contingencies, 2019–21', in J Blaxland (ed.), *On Mobilisation* (Cambridge University Press).

74 DoD (2019), *ADFP 5.0.1 Joint Military Appreciation Process*, 2nd edn, AL3 (Department of Defence, Australian Government).

75 Executive interviewed for this case study, June 2022.

76 There is wide scope for a range of interpretations of ‘discretionary’ tasking, e.g. duration, magnitude, geographic extent, concurrence.

77 There is a nuance here. Unlike a call-out (e.g. used for the national bushfire emergency), a call for the Reserve does not involve compulsion. However, during the pandemic, the majority of Reserve formations and units ‘turned off’ combat training and exercises. Therefore, from an individual Reservist perspective, the choice set was reduced, as there were few alternative options to serve in their part-time capacity. The generalisable scalability insight here is that scalability leadership can encourage staff to ‘choose’ to contribute to the emergent scaling requirement, by reducing alternative options in their existing or preferred work areas. This is at the heart of a scaling organisation’s determination of ‘essential’ vice ‘discretionary’ organisational activities during a scaling event.

78 McChrystal et al., 2015.

79 Economic inefficiency is distinguished here from administrative (or procedural) inefficiency. The former is desirable from a scalability perspective, if it is a consequence of securing latent capacity. In contrast, administrative inefficiency is undesirable from a scalability perspective, as it slows enterprise operational tempo. See Principle 5.

80 DB Copland (1942), *Towards Total War* (Angus and Robertson), p. 2.

81 McChrystal et al. 2015, p. 132.

82 ‘Catalyst: Chemistry. a substance that causes or accelerates a chemical reaction without itself being affected’, dictionary.com, https://www.dictionary.com/browse/catalyst, accessed 25 September 2022.

83 TB Crespi, P Rezende da Costa, T Scariot Preusler and CB Silva Cirani (2022), ‘Absorptive capacity in a public research company: from maturity to scalability’, *Brazilian Business Review*, 19(2), 133–152, https://doi.org/10.15728/bbr.2021.19.2.2

84 http://classic.austlii.edu.au/au/legis/cth/consol_act/da190356, accessed 1 November 2022.

85 https://www.pmc.gov.au/resource-centre/national-security/australian-government-crisis-management-framework, accessed 27 October 2022.

86 DoD (2021), *Defence Strategic Workforce Plan 2021–2040*, ‘Synopsis’ (Department of Defence, Australian Government).

87 http://classic.austlii.edu.au/au/legis/cth/consol_act/drsa2001340, accessed 1 November 2022.

88 http://drnet/vcdf/fd/fa/defence-mobilisation-planning/pages/default.aspx, accessed 31 October 2022.

89 See Chapter 3.

90 ADF 2024, p. 11.

91 Australian Army (2021), *Army Business Plan 2021–25* (Department of Defence, Australian Government), para. 1.16, p. 6.

92 DoD (2022), *2022–26 Defence Corporate Plan* (Department of Defence, Australian Government).

93 DoD (2021), *Defence Annual Report, 2020–21* (Department of Defence, Australian Government).

94 McChrystal et al. 2015, p. 32.

95 JB Barney (1991), ‘Firm resources and sustained competitive advantage’, *Journal of Management*, 17, 99–120.

96 M Scott (2022), ‘Many strategists but little strategy: Australia’s military strategy absence’, *Australian Journal of Defence and Strategic Studies*, 4(1), 39–64, https://www.defence.gov.au/sites/default/files/research-publication/2022/AJDSS-V4N1-web-full.pdf

97 Ibid.

98 ADF 2024.

99 DoD (1987), *The Defence of Australia* (Department of Defence, Australian Government), https://www.defence.gov.au/sites/default/files/2021-08/wpaper1987.pdf, accessed 1 November 2022.

100 DoD (2021), *AUKUS: Trilateral Security Partnership*, Fact Sheet (Department of Defence, Australian Government), ParlInfo - Fact sheet: Trilateral Australia-UK-US Partnership on Nuclear-Powered Submarines (aph.gov.au)

101 DoD (2016), *Integrated Investment Program* (Department of Defence, Australian Government).

102 ADF 2024, p. 17.

103 See Chapter 3.

104 Defence Force Recruiting (2021), *Defence Force Recruiting Business Plan 2021–2022* (Department of Defence, Australian Government). See also Defence Force Recruiting (2021), *Defence Force Recruiting Strategic Plan 2021–26* (Department of Defence, Australian Government).

105 DoD (2020), *Lead the Way: Defence Transformation Strategy* (Department of Defence, Australian Government).

106 E.g. Harmon 2014, p. 266: ‘It is rarely cost-effective to try to automate the work of a human expert. As expensive as it is to maintain such experts, it is cheaper to hire them and pay them to remain up to date than to try to capture and automate their knowledge.’

107 RH Thaler and CR Sunstein (2009), *Nudge* (Penguin).

108 Australian Government (2023), National Defence: Defence Strategic Review (Commonwealth of Australia), https://www.defence.gov.au/about/reviews-inquiries/defence-strategic-review

109 UAV = uncrewed aerial vehicle.

110 H Shams, ‘NSW inquiry into ambulance ramping told patients “dying unnecessarily”’, ABC News, 5 October, https://www.abc.net.au/news/2022-10-05/nsw-patients-dying-unnecessarily-ambulance-inquiry-told/101504524, accessed 1 November 2022.

researchcentre.army.gov.au