Improving IT effectiveness & efficiency

Service Lifecycle Management for Virtual Infrastructure

4sl Consulting Group 80-83 Long Lane, London EC1A 9ET
t:: 0207 692 7017 **w:** www.4slconsulting.com

1. Why read this?

1.1. Are you getting the most out of your Virtual Infrastructure?

The era of distributed systems and client server technology has left a legacy of diverse infrastructure components and complex configurations. Many data centres are littered with half-forgotten remnants of past IT programmes. Silo behaviour has yielded underutilised hardware: it is not uncommon to find costly capital assets sitting around unused for 70% or more of their time

Many organisations have recognised that the traditional approach to satisfying compute requirements with dedicated infrastructure is simply untenable. In response to this, server rationalisation and consolidation initiatives have been undertaken to deliver increased levels of utilisation across the existing and future server environment. Against this backdrop, virtualisation technology has matured to the point that many enterprises are accelerating their consolidation plans with server and desktop virtualisation to realise greater benefits.

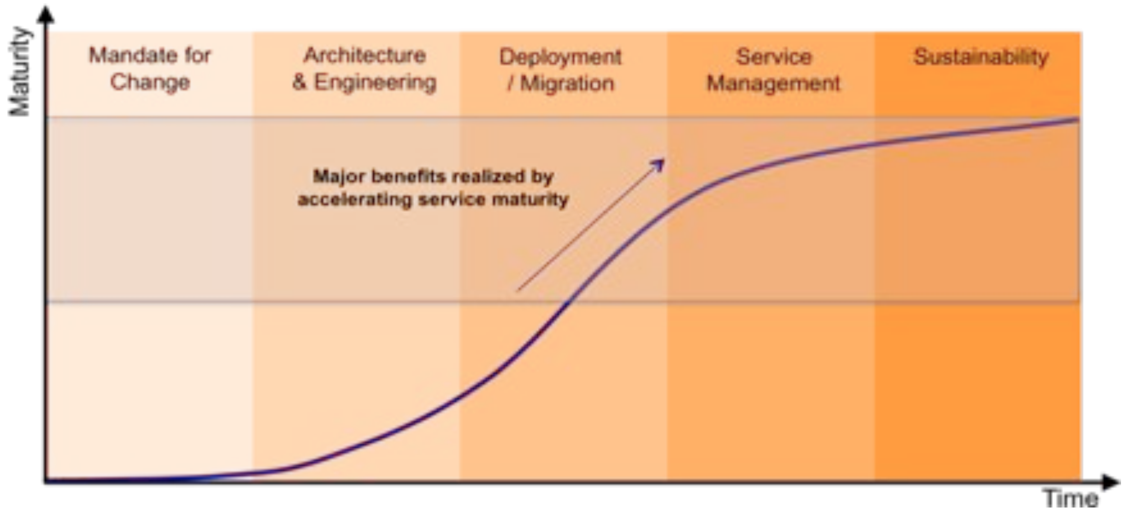
However, a number of organisations are not necessarily seeing the full range of benefits they expected, or at least some benefits come at a cost elsewhere. For example, an increase in operational expenditure to manage, support and sustain an enterprise-wide solution may in some cases outweigh the initial capital expenditure savings. Other expected improvements in operational efficiency such as faster provisioning time, greater resource recycling and management may have also failed to live up to expectations.

This short guide aims to identify some potential obstacles to the full realisation of the benefits of a virtual infrastructure and how, with proactive Service Lifecycle Management, they can be overcome.

1.2. What is Service Lifecycle Management?

We recognise that any new IT service offering will be subject to some formal, or more often, informal and adhoc periods of activity from initial conception to full deployment and operational support. However, not all of the activity is seamless and there are often clunky and painful interactions (or handovers) along the way. This can impact the efficiency and effectiveness of the service itself increasing the time to deployment, functionality and availability and the bottom line cost.

In response to this we have mapped out a range of best practices, desired behaviours and disciplines against a preferred set of outcomes - collectively we call this Service Lifecycle Management. At 4sl, our primary aim is to help clients increase IT service efficiency and effectiveness we underpin our approach with the model as shown in the diagram below.



This paper applies the lifecycle model to the development and delivery of an enterprise class virtualisation infrastructure, highlighting the major milestones and deliverables associated with it.

Whilst it is not meant to be prescriptive it does aim to demonstrate a repeatable and proven method to move seamlessly from an initial assessment, through business case preparation and proof of concept to full implementation and benefits realisation.

2. Where to Start?

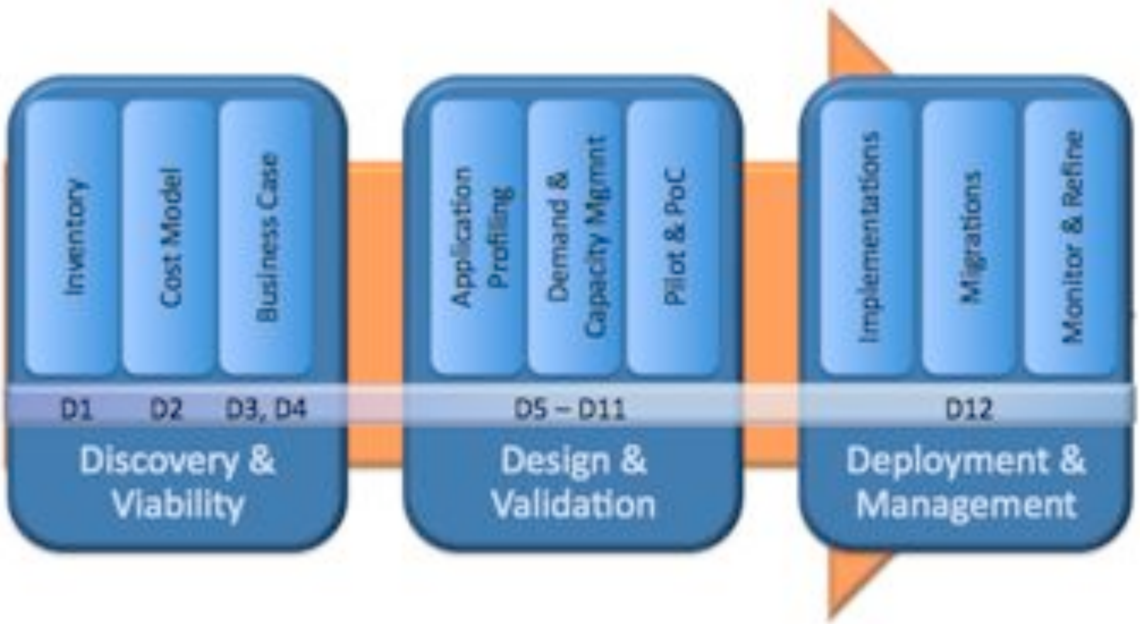
Often at the beginning as the lifecycle model suggests, however, that's not always possible. System and services evolve, they undergo upgrades, migrations, get expanded and refined. They are often subject to major 'in-flight' change (or operational engineering as we'd like to see it). In any event, the desire to realise grater benefits from an IT environment often requires invasive activity with operational services.

The diagram below illustrates the main phases of initial activity employed by 4SL on a typical server virtualisation initiative, namely:

- Discovery and viability. The initial phase will typically use inventory information on the server estate to help develop a financial model to identify the costs involved and potential savings that could be realised from a virtualisation programme. A situational assessment is also executed to determine any constraints, issues or potential changes required to the operational environment in order to manage a virtualised infrastructure. The results from both initiatives would be presented through a formal business case for review and approval.

- Design and validation. Often undertaken as an engineering activity to develop and validate an architecture of the desired environment but with consideration of the target application sets, future demand and other service characteristics. The design will be validated through a Proof of Concept or Pilot implementation.

- Deployment & Management. Formal implementation and/or migration to the design virtualised environment. Typically run as project but with service delivery & support functions fully engaged to ensure a seamless handover to operational support.



Each phase includes specific service offerings that 4SL can deliver either in a stand-alone manner or as programme of consecutive phases. Each phase has a corresponding set of deliverables – these are numbered and referred to in more detail in the sections that follow.

3. Virtualisation Viability Assessment

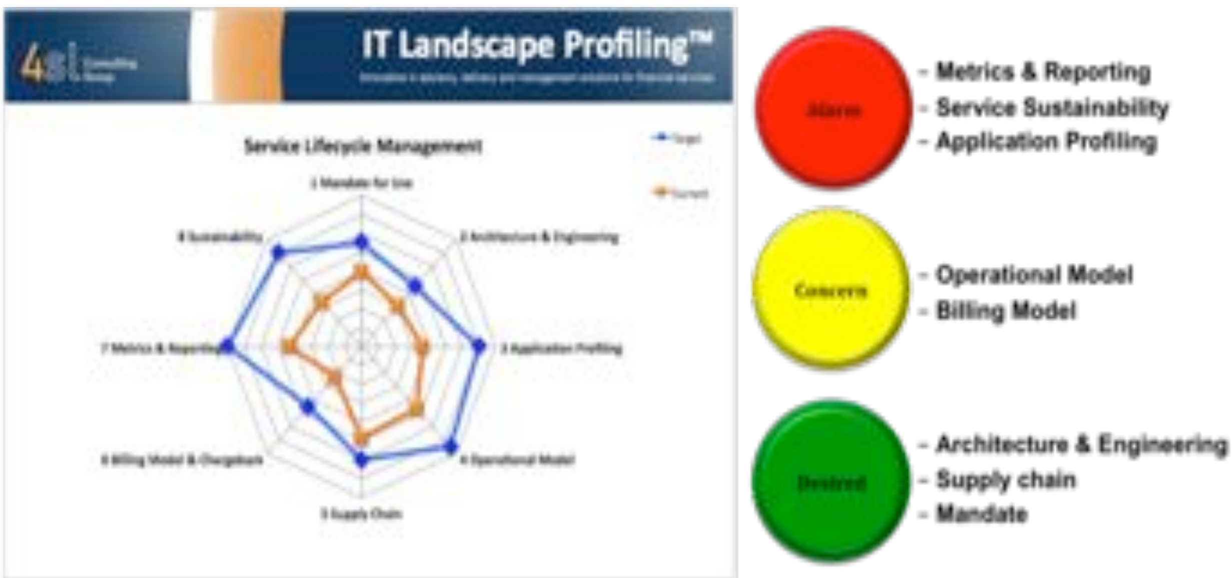
This is typically a sub 2-week engagement where virtualisation objectives, plans and any associated activities and practices are assessed in the context of the 4SL service lifecycle.

3.1. Rapid Situational Assessment

4SL recommends that this rapid assessment is undertaken as the first step in any server consolidation or rationalisation initiative before an organisation embarks on potentially costly consulting or implementation pilots. The viability assessment includes:

- A kick-off stakeholder workshop to define scope, drivers, objectives, current pain points, input data and output deliverables.
- The collection of infrastructure system details (typically directly from an inventory) along with a number of performance and workload metrics from a targeted section of the current infrastructure estate.
- A high level review of existing service delivery & support practices where a gap analysis is performed to determine the delta between the desired (or target) state and the current state.

4sl have developed a rapid assessment technique called IT Landscape Profiling™ which uses subject matter expertise across a number of virtualisation technologies and draws on experience gained from a range of previous engagements. Eight functional areas are given a priority weighting and then assessed through a structured Q&A process. The output typically takes the form of the graphic as shown below:



There are two major deliverables resulting from the viability assessment:

- **D1. Viability Report –** a document that details potential options, cost model and benefits for virtualisation adoption, based on the stakeholder requirements and objectives. It will include the results from the Landscape Profiling process by highlighting potential gaps, issues and constraints. A summary of the inventory report is also included along with a number of recommendations for service improvement, operational efficiency and risk mitigation measures.
- **D2. High-level design** – a document that presents an outline architecture for the target 'virtualised' infrastructure and forms the basis of the subsequent design phase.

3.2. Application Profiling, Metrics & Collection

A mandatory step in the development of a successful business case.

The baseline reference against which performance and availability are measured.

The collection of system metrics is a key process in any server consolidation or rationalisation initiative. 4SL sees this as a mandatory step in the development of a successful business case. It also provides a baseline reference against which performance and availability can be measured post virtualisation. The process is underpinned by a reliable and reusable set of principles to enable the transition from the physical infrastructure estate to the target 'virtual' environment.

For the purposes of the viability clinic engagement, a sample of the overall server estate is selected along with the details of the applications that run across the target group. The key metrics of computing activity that need to be analysed from this sample are:

- Processing Capability
- Memory
- Disk
- Network

It is recognised that each attribute cannot be assessed in isolation. However, a common goal is to adopt a simple, repeatable and proven model whilst recognising that through the physical to virtual transition (called P2V), a margin of error exists and scope for growth will be required. It is also recommended that the performance metrics of the physical infrastructure are graphed to determine 'peaks', i.e. when a system's workload exceeds its average utilisation for a sustained period.

3.2.1. Processing Capability

The following metrics are recorded (from the perspective of what the operating system sees as its utilisation, i.e. against all processors that are being utilised):

- total processor average utilisation over the analysis period
- percentage of time a system is running above 90% total processor utilisation

It is recognised that although a physical server may include multiple processors, the operating system and applications may or may not be effective at utilising all of them. To qualify the results, the total processing capability is noted. This is based on the number of physical processors, availability of hyper-threading or dual core processors and the clock speed of each processor.

One of the major benefits of virtualisation can manifest itself at this stage, i.e. identifying groups of applications that demand processing resource, but do not typically peak at the same time. Consolidation at this level ensures the overall computing estate can be optimised for greater efficiency.

3.2.2. Memory

Starvation of memory can have a dramatic effect on the computing performance, primarily due to memory pages being swapped to disk. The same constraint applies in the virtualised environment. Although it is still highly desirable to avoid swapping, a compromise is required in order to cater for peak bursts in memory utilisation. Also, in the virtual world, memory can be allocated and removed dynamically, as and when required. To qualify the results, the total amount of memory allocated to the server is noted.

The following metrics are recorded (this is from the perspective of what the operating system sees as its utilisation):

- total memory average utilisation over the analysis period
- percentage of time the system is engaged in page swap activity

3.2.3. Disk

There are two elements of disk resource that can be tracked; throughput and capacity. However, for the purpose of virtualisation, the most important factor is I/O through the disk controller. Again, the transition from P2V could result in a difference between the source technology (e.g. IDE based disk) and the target (e.g. Fibre Channel SAN storage). However, the main objective is to identify intensive I/O activity.

The following metrics are recorded (this is from the perspective of what the operating system sees as its disk utilisation[1]):

- average disk utilisation over the analysis period
- percentage of time the system is queuing for disk resource.
- the total amount of disk space available
- amount of disk space utilised

3.2.4. Network

Many systems have multiple network cards, typically these are consolidated (or 'teamed') to increase resilience and/or performance. The total available network resource is recorded through the following metrics:

- percentage of time the total network is running over 90% bandwidth utilisation
- total number of network cards,
- bandwidth of each card
- duplex settings of each card

[1]. It should be noted that disk performance counters are usually disabled as they can have an impact on system performance. It is at the discretion of the client to exclude disk analysis if deemed appropriate.

3.2.5. Summary of Infrastructure Metrics

Resource	Metric	Notes
Processor	Quantity	All four recorded to evaluate the available processor capability.
Processor	Hyper Threading	
Processor	Dual-core	
Processor	Clock Speed	
Processor	Average Utilisation	
Processor	Time spent swapping	
Memory	Quantity	
Memory	Average Utilisation	
Memory	Time swapping	Assumes memory utilisation of 100%
Disk	Total Size	
Disk	Space Utilisation	
Disk	Average Utilisation	
Disk	Time queuing	Assumes disk I/O utilisation is 100%
Network	Quantity	
Network	Speed	
Network	Duplex	
Network	Average Utilisation	

3.3. Example of potential consolidation ratios

As indicated earlier, the critical mass of system telemetry metrics that are recorded during this phase form the basis upon which consolidation principles can be applied to determine the likely target, virtualised environment. The table below, prepared for a previous engagement, demonstrates the target consolidation ratio that can be expected against the levels of utilisation recorded across an estate of over 3000 servers.

3150 servers in total		
Category	Current utilisation	No of servers (Notes)
Core Infra Machines		648 (not hit in phase 1)
Stressed machines	Peak>75%, Average>75%	51
Well used servers	Peak>50%, Average>40%	447 (not candidates for consolidation)
Bursty servers	Peak>50%, Average<40%	390 (unsuitable because consolidation may constrain the peak)
50% utilisation	Peak 30-50%, Average<50%	201 (assume 2:1 consolidation)
30% utilisation	Peak 20-30%, Average<30%	165 (assume 3:1 consolidation)
25% utilisation	Peak<20%, Average<20%	1057 (assume 4:1 consolidation)

With very conservative consolidation ratios and avoiding any contentious mixture of workloads or applications, consolidating these machines would almost double average utilisation from 14% to 24% (see table below).

	No of Servers	Potentially spare machines
Core Infra Machines	x648	x150
Stressed machines	x51	-
Well used servers	x447	-
Bursty servers	x390	-
50% utilisation	x201	x101
30% utilisation	x165	x110
25% utilisation	x1057	x793
Un-used	x190	x143
Total	**x3149**	**x1296**
Utilisation	**13.9%**	**23.6%**

4. Business Case Development

Material delivered through the viability assessment is often used to help to prepare an internal business case for the adoption of server virtualisation. As highlighted in the service lifecycle framework, one of the key outputs from this initial phase of work is to gain a mandate from senior management for the adoption of virtualisation services. Without this the initiative will fail.

From previous client engagements, using the application profiling and metrics gathering process above, 4SL has gained significant experience in preparing business cases and recognises the typical upfront costs and associated benefits that can be realised.

Examples of objectives typically associated with a server virtualisation business case include:

- **TCO reduction** – i.e. lower overall cost of the Windows and Linux server (or desktop) estate by X%. This refers to the cost of providing and maintaining the hardware, and associated datacentre facilities. It does not include the people costs of supporting an OS image, or the application software running in the image. These are typically unaffected by the virtualisation program.
- **Time to market** – i.e. reduce the time taken to provision a compute instance by Y% from final approval to the system being made available and ready for use. In a virtualised environment, this benefit is typically realised through identification and delivery of unused compute capacity.
- **Service level improvement** - increase system availability through enhanced BCP solutions.
- **Operational efficiency gains** – i.e. reduce the complexity of the distributed systems environment by standardising on a common compute platform and enabling central management for the majority of applications and operating platforms.

The deliverables from this phase include:

- **D3**. Material to support the development of an internal business case.
- **D4**. Detailed cost/benefits analysis and ROI with some of the major costs groups identified to help develop a billing policy and financial chargeback model.

5. Design & implementation

For a full virtualisation implementation, 4SL would propose a formal project engagement with corresponding control, tracking and reporting functions provided by a dedicated project manager.

This phase will aim to leverage the high level design produced as part of the viability assessment (if it has been conducted) or review any design plans already prepared.

Over and above the disciplines and functions delivered through the 4SL service lifecycle methodology, the main deliverables produced during these phases of the project include:

- **D5. Design Specification(s)** – i.e. a collection of detailed design documents, architecture diagrams and plans used to systematically build a specific virtual infrastructure.
- **D6. System & Acceptance Test Plan(s)** – i.e. a collection of test plans designed to validate the functionality and performance of system components such as dynamic resource allocation, physical server to virtual machine migration (P2V), and acceptance of operational functions such as failover and backup, as well as high-level benefits realisation of server consolidation ratios.
- **D7. Overall Project Plan** – a detailed project plan that includes major phases, milestones and deliverables aligned against resources, budgets and anticipated infrastructure availability.
- **D8. Working infrastructure** – the primary deliverable is a working virtual infrastructure built to the agreed design. This could be developed from an initial Proof of Concept, through a Pilot deployment to full enterprise adoption. This will depend on the organisation's promote to production strategy for such infrastructure change.
- **D9. Certification & Acceptance Test Report** – full and detailed results of executing the steps outlined in the Test Plan(s).
- **D10. Exception Report** – highlights differences or gaps between the plan and actual implementation.
- **D11. Finalised Design Specification(s)** – A revision of the detailed design that incorporates lessons learned from the actual build process.

6. Operational Readiness

Assessing the impact on service delivery and support resources.

Highlighting areas for service change or improvement.

The impact of introducing new virtualisation techniques is likely to have a major effect in the service delivery and support functions.

4SL has significant experience in defining and delivering the operational transition process for new infrastructure services. The 4SL service lifecycle introduced earlier contains a formal service management phase that promotes the best practices and procedures from the ITIL framework. In this context, 4SL recognises that the traditional approach to deploying compute resources will need to be adapted in order to realise benefits such as a faster provisioning cycle. For example; configuration, change, availability, demand and capacity management will need to be analysed - significant gaps between the current processes and procedures and those required for a virtualised environment may need to be addressed. Another area likely to be impacted is chargeback and financial management. In the traditional siloed environment there is typically a hard-wired relationship between IT infrastructure assets, an application and a given line of business (LOB). This enables services associated with those assets to be allocated directly to a given business unit or cost centre. This may need to be reassessed to accommodate the virtualised world.

As part of the initial virtualisation viability clinic, a high-level impact assessment of the service management functions will be carried out. Although this will not provide a full assessment of operational readiness, it may highlight areas for service change or improvement that that should be addressed during the Design & Implementation phases.

Dynamic workload management

In the virtualised world, there is now the ability to define service levels for the processor on a machine-by-machine basis. Also, if resource capacity is a concern, it is possible to set alert thresholds and move the demanding application or workload around the compute farm. This dynamic nature of workload management may not be desirable over a longer period, but it does provide flexibility to handle short term and well-defined periods of demand and can be adopted to satisfy business service levels with minimal impact on system operation or cost.

Interim operations management

An organisation may require a period of interim operations management whilst the range of virtualised services is deployed into the production environment. The major deliverable to enable handover to the business as usual (BAU) teams is:

- **D12. Operational Support Guide** – a comprehensive set of operational and support guides (or ‘run books’) to enable service delivery teams to manage the target virtual infrastructure