VIEW KEY INSIGHTS (HTTP://WWW.GARTNER.COM/DOCUMENT/2828319/KEY-INSIGHT?REF=DDISP)

How to Build Best-Practice Infrastructure Capacity Plans

21 August 2014 G00268643

Analyst(s): *Ian Head* | *Milind Govekar*

Summary

A key ITSM discipline, capacity and performance planning is key to delivering infrastructure that is agile and capital-efficient. Infrastructure leaders can use such approaches to proactively justify investment, reduce the risk of SLA breaches due to shortfalls and increase business satisfaction.

Overview

Key Challenges

Ensuring demonstrably cost-effective IT infrastructures is impractical without a history of regular, good-quality capacity plans.

Producing effective overall infrastructure investment forecasts is difficult when capacity plans are produced in technical siloes, with inconsistent formats, assumptions and timescales.

Most organizations try to keep excessive capacity headroom (just in case), which wastes money and power, and produces wasteful emissions.

Outages and poor performance due to capacity shortfalls still commonly provoke urgent, premium-priced upgrades.

Recommendations

I&O leaders should mandate infrastructure managers to use the format provided in this best practice research to produce regular capacity plans. This will improve service quality, business satisfaction and investment effectiveness.

Begin with capacity planning of the technical infrastructure (component capacity planning), but then drive forward into service and business capacity planning for increased business alignment and prestige with business customers.

Use this guidance to produce infrastructure capacity plans that meet industry best practices, including ITIL.

Introduction

This research provides clear guidance on the contents to be expected of best-practice IT infrastructure capacity plans. Capacity plans are used to justify investment capital and reduce the risk of outages and performance degradation. These traditional reasons for capacity planning have been joined by a newer, more pressing demand.

The reliability expectations for IT systems keep increasing, but so does the need for agility. The increased use of agile methodologies such as lean and DevOps demonstrate that improved agility, more rapid time to market, higher performance and more efficient use of investment capital have all risen in importance in our clients' priorities. [1] (#dv_1_over_200) , [2] (#dv_2_executive_summary) , [3] (#dv_3_executive_summary) To achieve a flexible infrastructure that meets these needs requires that infrastructure managers produce capacity and performance plans with clear, succinct, relevant information to allow I&O leaders to allocate resource and capital where it will deliver the most business benefit. Gartner's experience is that capacity plans often lack key information, which adds unnecessary risk to the decisions made. This research states the content that infrastructure managers should include in the capacity and performance plans in an organization that has achieved or is aiming to achieve Level 3 or better in Gartner's ITScore for Infrastructure and Operations (ITSIO) methodology. These recommendations align with ITIL.

This research should be read in conjunction with:

"Govern the Infrastructure Capacity and Performance Planning Process With These 13 Key Tasks" (http://www.gartner.com/document/code/268641?ref=grbody&refval=2828319)

"How to Create and Manage an Infrastructure Capacity and Performance Plan" (http://www.gartner.com/document/code/268640?ref=grbody&refval=2828319)

Analysis

Physical and Virtualized Resources Need Capacity Planning

As the IT infrastructure environment evolved away from mainframe-centric architecture, capacity planning of the entire IT infrastructure has tended to decrease in priority, especially in the past 15 years. The approach taken has often been to keep a large amount of capacity headroom, since the costs associated with this were not seen as excessive, outside the mainframe environment. Additionally, the bandwidth requirements of newer technologies, such as voice over IP (VoIP), unified communications, desktop video and the emergent Internet of Things, require capacity and performance planning that goes far beyond the server estate.

As CIOs survey their current estates, still with a large number of physical servers, but also with ballooning virtual environments, and storage and network requirements rising much faster than Moore's law, it is clear that more control needs to be exercised. The need for capacity and performance planning as a discipline, supported by a regularly updated capacity and performance plan, has once again become obvious and urgent in many organizations.

Enterprises with highly virtualized environments may apply more focus on optimizing the resource allocation and performance within a given environment, and drive toward a capacity-on-demand philosophy. However, even in these circumstances, additional infrastructure is neither instantaneous nor free, and so capacity and performance planning remains a key skill in delivering an infrastructure that is both capital-efficient and optimized for the different services delivered — whether they are commercial off-the-shelf (COTS) or developed using agile or waterfall methodologies.

Create a Hierarchy of Capacity and Performance Plans

Organizations typically divide responsibility for their infrastructure into discrete segments under the responsibility of technology towers, IT service managers, service providers and possibly business units. Therefore, the I&O leader or CTO should require a set of plans for subsections of the infrastructure, which are then ultimately combined into an overarching capacity and performance plan for the entire infrastructure under his or her control. Depending on the analytics tools available, drill-down and additional analysis may also be possible to assist the investment decisions further.

The plan aggregation procedure can be relatively simple if each plan in the hierarchy conforms to the same structure and format. Organizations at ITScore maturity Levels 1 and 2 will usually produce plans at the component level (for example, Intel data center estate, WAN and Oracle licenses), while the plans of those organizations at maturity Level 3 and above will also contain capacity and performance plans aligned with specific service and business processes. The latter plans are much more likely to reflect direct input from the business and, therefore, be much more influenced by projected business plans and activities, rather than rely on historical trend analysis. For more information on using ITScore to guide service improvement activity, see "What You Need to Know About ITSIO." (http://www.gartner.com/document/code/250747?ref=grbody&refval=2828319)

It is, therefore, clear that the detailed content of the sections described below will vary, depending not only on the details of the technology infrastructure, but also on the maturity of the organization. When using this guidance to set standards within your organization, include specific guidance on the component, service and business content expected in the plans to be created in your organization.

Include This Content in the Outline of Each Capacity and Performance Plan

The guidance below reflects Gartner's view of best-practice capacity planning and also aligns with ITIL. All the following sections should be present in each capacity and performance plan.

DOCUMENT CONTROL

The set of capacity plans should be part of the organization's document control methodology and should, therefore, contain history, versioning, ownership and relationship information. The set of plans will also be part of the capacity management information system where this has been implemented.

EXECUTIVE SUMMARY

A prime goal of the document is to provoke investment decisions for the right areas and to avoid or postpone investment decisions where appropriate. Therefore, this section should cut straight to the main business issues to be addressed, and decisions needed — remembering that a decision to take no action is still a key decision.

Avoid any technology detail not essential to the decision. State the issues of most business impact, the costed options and briefly justified recommendations.

The overall infrastructure capacity and performance plan is likely to contain many recommended investments and complex financial analysis, so the executive summary of the overall infrastructure capacity and performance plan may be spun off into a separate document and focused on specific financial and business-based readerships.

INTRODUCTION AND SCOPE

State which segment of the infrastructure this plan pertains to. The overall infrastructure plan should cover the entire estate; but where this specific plan is intended as part of a set that is aggregated into a larger plan, it should be stated. In some organizations, the scope of the capacity plan also includes the facilities, people resources and skills required to plan and manage the infrastructure.

Therefore, this section should contain summary information on:

- The time scope of the plan — that is, whether this is an annual, six-month or rolling monthly plan
- The components, services, facilities, resources and skills within the scope of this plan
- The current levels of capacity
- The current performance delivered, including service-level achievement and information on the performance incidents logged
- A summary of incidents caused by undercapacity
- A view on when service incidents or financial impact are envisioned due to over or undercapacity
- Changes in the infrastructure, business environment, plans and forecasts since the last issue of the plan (for example, corporate acquisitions, mergers or downsizing plans)

BUSINESS ENVIRONMENT AND BUSINESS SCENARIOS

The plan should include the potential infrastructure impact of new application and network services, in addition to services scheduled for reduction or withdrawal.

The plan should state all the forecasts and scenarios that have been used in compiling the plan, and the options for fulfillment so that the IT and business leaders may make judgments on the likelihood and impact of each plan.

In enterprises at low levels of business or IT maturity, the infrastructure team may have very little business information on which to base the capacity and performance plans and, therefore, may be working almost entirely from historical data. If this is the case, then this should be stated in the executive summary, as well as in this section, because there are clear risks associated with this approach. A plan that assumes that the future is predictable purely from past data clearly carries more risk than a plan with strong input from business forecasts and plans. Mitigating the associated risks is likely to mean either higher excess capacity or increased risk of performance and capacity shortfalls — or both.

When the capacity planner is able to work closely with the business, then the plan can reflect the planned business targets and activity, and also contain contingencies (for example, the expected impact of a promotion may be a 5% increase in monthly transactions, but if the increase exceeds this value, then provision may be made for relocating workloads and reallocating storage to cope with the success, up to defined limits). This is an example of proper planning preventing poor performance.

METHODS USED

Capacity and performance management is highly dependent on information provided by other processes, and the establishment of a baseline. This section should state the sources of information, the tools used to gather and analyze information, and the methods used to model the impact on the infrastructure and service performance.

This is likely to include monitoring data from infrastructure components, application performance monitoring (APM) tools, network

performance monitoring and diagnostics (NPMD) tools, business forecasts (including macroeconomic impacts), workload forecasts, modeling techniques, and the output from service modeling tools.

This should include performance, availability and service-level data normally produced by existing monitoring tools that analyze infrastructure components and application execution, producing detailed workload forecasts and statistics on end-user delivery. More-mature organizations will also include business forecasts (including macroeconomic impacts), modeling techniques used and the output from service modeling tools.

ASSUMPTIONS MADE

State all business, economic and technical assumptions made in the production of the plan. This is quite difficult, since the most dangerous assumptions are those that we don't consciously make — witness the historic difficulties in the eurozone and volatility in macroeconomic conditions.

Be as rigorous as possible in stating the assumptions that underpin the stated plan and decisions requested.

Modeling errors are less common than failures in the assumptions concerning business drivers and forecasts.

SERVICE DEMAND SUMMARY

Where the organization is mature enough to report in terms of IT services, this section should be included and should profile the IT services provided in terms familiar to the service managers and business leaders — for example, service transaction peaks, mean and total number of accounts processed, and so on. Profiles and forecasts should be provided for new and existing services, including any plans for service retirement.

Short-term, midterm and long-term forecasts should be included based on the best information available from business plans, promotion and activity schedules.

Critical and high business impact services should be profiled individually, while less-critical services may be aggregated, provided this still permits the resource demand to be estimated.

RESOURCE DEMAND SUMMARY

This section translates the forecast service demand into utilization of infrastructure resources — for example, processor, memory, storage, licenses, bandwidth, data center and power. Best practice is to use service capacity models to make these calculations. Investment in capacity tools can be highly effective in improving the quality of the resource demand analysis and the visual impact of the material.

Gathering the data necessary to write this section will entail close working with the infrastructure and monitoring technical staff. Close cooperation in this effort among the various technical teams will often yield insights in how to make better use of the existing investment to obtain better data and better performance.

This section should show the recent utilization across all the infrastructure towers, and also forecast resource demand at least 12 months ahead. Any critical dates where infrastructure constraints would produce performance incidents or outages should be highlighted. The Options for Service Improvement and Recommendations sections should contain suggestions for avoidance of the negative business impact.

OPTIONS FOR SERVICE IMPROVEMENT

There are nearly always alternative ways of relieving capacity shortfalls and improving performance. This section should detail the options considered, together with the pros and cons, and which scenarios each option is best-suited to. The options should not include only acquiring new technology. Service improvements may involve people, processes, management and technology innovations.

This section should be written using the knowledge of the architecture standards and strategic technology policies in place in your organization. This should include:

- Options that include optimizing the use of existing resources, sharing infrastructure and services, and increasing standardization and automation.

- Replacing or rewriting legacy systems to achieve compatibility with new policies and platforms, including server virtualization, software-defined networking (SDN), network function virtualization (NFV) and cloud service providers.

- The corporate social responsibility and financial benefits of reduced power and carbon emissions. Total cost of ownership should be

considered when considering whether to replace older equipment or upgrade hardware and software.

COSTS FORECASTS

The cost of each option, and practical combinations of options, should be stated here. The overall infrastructure capacity plan should also include the current and forecast costs of providing the totality of IT services to the business.

This is an important integration point with financial management and business relationship management. Capacity management needs budget and other information as input, while financial management needs investment demand forecasts associated with potential future business scenarios; business relationship management needs to know the investment required to deliver the business plan.

RECOMMENDATIONS

This section should refer back to previous versions of the plan, and state which of the earlier recommendations have been rejected, planned and implemented. Variances between the expected and actual costs and impact should also be noted.

The recommended options should be justified, including potential consequences if the recommendations are not adopted.

Each recommendation should state:

- Business benefits to be expected, including time scales
- Business and technology impact of the recommendations
- Resources required
- Costs — both one-off and ongoing
- Risks

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"Well-Defined Duties of the Process Owner and Process Manager Are Critical Success Factors for Service Improvement Programs" (http://www.gartner.com/document/code/259405?ref=ggrec&refval=2828319)

"Govern the Infrastructure Capacity and Performance Planning Process With These 13 Key Tasks" (http://www.gartner.com/document/code/268641?ref=ggrec&refval=2828319)

"How to Create and Manage an Infrastructure Capacity and Performance Plan" (http://www.gartner.com/document/code/268640?ref=ggrec&refval=2828319)

"Adding Business Value Through Enhancing Capacity Management Maturity" (http://www.gartner.com/document/code/218990?ref=ggrec&refval=2828319)

"Attaining Capacity Management Process Level 2: Establish Solid Capacity Planning Basics" (http://www.gartner.com/document/code/214905?ref=ggrec&refval=2828319)

"Attaining Capacity Management Process Level 3: Introduce a Service Focus and Integrate With Demand Management" (http://www.gartner.com/document/code/214906?ref=ggrec&refval=2828319)

"Attaining Capacity Management Process Level 4: Extend the Service Focus to Integrate With Business Strategies" (http://www.gartner.com/document/code/218971?ref=ggrec&refval=2828319)

"IT Infrastructure and Operations: Still Immature After All These Years" (http://www.gartner.com/document/code/211762?ref=ggrec&refval=2828319)

"Capacity and Performance Management Form the Basis of Web-Scale IT" (http://www.gartner.com/document/code/261096?ref=ggrec&

refval=2828319)

"How Antifragile Practices Can Make Your I&O Stronger" (http://www.gartner.com/document/code/261095?ref=ggrec&refval=2828319)

"How to Determine Readiness for Voice, Video and Unified Communications" (http://www.gartner.com/document/code/247761?ref=ggrec&refval=2828319)

"Use These Best Practices to Manage IP Voice, Video and Unified Communications Deployments" (http://www.gartner.com/document/code/248201?ref=ggrec&refval=2828319)

Evidence

[1] Over 200 Gartner-client engagements, interactions and inquiries every year.

[2] "Executive Summary: Reimagining IT: The 2011 CIO Agenda" (http://www.gartner.com/document/code/210382?ref=grbody&refval=2828319)

[3] "Executive Summary: Leading in Times of Transition: The 2010 CIO Agenda" (http://www.gartner.com/document/code/173968?ref=grbody&refval=2828319)

C. Rudd and V. Lloyd, "Service Design Book," The Stationery Office, 2007.

J. Allspaw, "The Art of Capacity Planning: Scaling Web Resources," Oreilly & Associates, 2008.

L. Klosterboer, "ITIL Capacity Management," IBM Press, 2011.

Document Revision History

How to Build Best-Practice Infrastructure Capacity Plans - 21 August 2014 (http://www.gartner.com/document/2828319?ref=ddrec)

How to Build Best-Practice Infrastructure Capacity Plans - 27 January 2012 (http://www.gartner.com/document/1907714?ref=ddrec)
