The Storage Capacity Design Dilemma
an ITIL approach

LeRoy Budnik
Knowledge Transfer
SNIA Legal Notice

• The material contained in this tutorial is copyrighted by the SNIA and portions are subject to other copyrights.1
• Member companies and individuals may use this material in presentations and literature under the following conditions:
  – Any slide or slides used must be reproduced without modification
  – The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
  – This specific legal notice shall not be removed.
• This presentation is a project of the SNIA Education Committee.
Abstract

The Storage Capacity Design Dilemma

As architects, we must continually discover capability limits, constraints, and patterns and match requirements, capabilities and cost to provide an effective design. When we add the business requirements, it becomes the core mantra of ILM. Yet translating these requirements into hardware and software is not easy. In the process, we must continuously choose between complex alternatives, some of which seem equally unacceptable, and hence the dilemma. In this session, we will introduce fundamentals of storage infrastructure design from the perspective of ILM requirements. We will introduce key formulas, the science of storage capacity planning and follow a case study to demonstrate application in host, fabric and array design, to meet the requirements. The dilemmas of design decisions clear up once you know the formulas. It will change your thinking.
Objectives

In this session, we have a dual focus on the science and consultative process of developing a mature discipline of storage capacity assessment, planning and design. The IT Information Library (ITIL) is the foundation of the methodology. Essential formulas come from a variety of sources.

Objectives:

• Know basic storage capacity and performance formulas
• Design for capacity and scalability in hosts, SAN, storage arrays and backup
• Understand how to break out of the most common storage design dilemmas
Capacity Management

Capacity Management ensures that IT process and storage capacity provisioning match evolving business demand in a cost effective and timely manner.

**Goal**
To understand the future business requirements, the organization's operations and IT infrastructure to ensure that all current and future capacity and performance aspects of the business requirements are provided cost effectively.

**Scope**
Business Strategy and Plan, IT Strategy and Plan including all applications, hardware, O/S, networking, peripherals and human resources.
Process

**Input**
- Technology
- SLAs, SLRs and Service Catalogue
- Business plans and strategy
- IT plans and strategy
- Business requirements and transaction volumes
- Operational schedules
- Deployment and development plans and programs
- Expected, scheduled Changes
- Incidents and Problems
- Service Reviews
- SLA breaches
- Financial plans
- Budgets

**Processes**

**Business Capacity Management**
- trend, forecast, mode, prototype, size and document future business requirements

**Service Capacity Management**
- monitor, analyze, tune and report on service performance, establish baselines and profiles of use of services, manage demand for services

**Resource Capacity Management**
- monitor, analyze, run and report on the utilization of components, establish baselines and profiles of use of components

**Output**
- Capacity Plan
- Capacity Database (CDB)
- Baselines and profiles
- Thresholds and alarms
- Capacity reports (regular, ad hoc and exception)
- SLA and SLR recommendations
- Costing and charging recommendations
- Proactive changes and service improvements
- Revised operational schedule
- Effectiveness reviews
- Audit reports
Iterative Capacity Management Activities

- **Tune**
  - Capacity
  - Implement
  - Analyze
  - Monitor
  - SLM Thresholds
  - Resource Utilization Thresholds

- **Capacity Management Database (CDB)**

The diagram illustrates the iterative process of capacity management activities, starting with implementing capacity, followed by monitoring, then analyzing, tuning, and finally implementing again, ensuring efficient resource utilization and capacity management.
Workload Management

- Produce Forecast
- Transform Aggregates into Hardware Forecasts
- Perform Workload Forecasting
- Classify Workload by Type
- Filter Necessary Metrics
- Create Workload Catalog
- Analyze Trends
- Develop Resource Profiles
- Capacity

Technical Metrics:
- Utilization
- I/Os
- Block Size
- Transfer Rates
- Cost

Process Metrics:
- MTTP
- MTTI
- MTTM
- MTBF/MTTR
Producing a Capacity Plan

- Business Capacity Management (BCM)
- Service Capacity Management (SCM)
- Resource Capacity Management (RCM/SRM)

**Iterative Activities**
- Demand Management
- Modeling
- Application Sizing

**Capacity**
- Produce Capacity Plan

**CDB**

Capacity Plan
Covers all aspects of BCM, SCM and RCM
Costs, benefits, problems
Planning and Implementation

Plan

Implement

Review

Synchronize

Capacity Plan

$ t $
Process Review

Review Goals
- Is required output available at the required times for the appropriate audience?
- Is its activity cost effective?

Validate Metrics
- Calculations and formulas
- Component utilization
- Acceptable data collection
- Accuracy of tuning predictions
- Are potential and actual breaches predicted and associated with full notification?
- Are imposed constraints within acceptable client limits?
- Is reporting regular and on time?
- Is the capacity plan produced on time?
- Are recommendations clear and accurate?

Critical Success Factors
- Accurate business forecasts
- Knowledge of IT strategy and plans, and that the plans are accurate
- Understanding of current and future technologies within the capacity management team
- Ability to demonstrate cost effectiveness
- Interaction with other Service Management processes
- Storage capacity matches business need

Key Performance Indicators
- Resource forecasts
- Technology
- Cost-effectiveness
- Measurement of accuracy of match between IT Storage Capacity and Business Need.
Capacity Process Interfaces

Service Level
- Requirements
- Targets, Achievements
- Incident

Availability
- Sufficient capacity reduces availability risks
- Provides cost summary in exchange for budget
- Configuration

Financial
- Continuity depends on capacity guarantee
- Change

IT Service Continuity
- Incident
- Problem
- Real time capacity monitoring resulting in incidents
- Supports analysis of performance trends
- CDB is a subset of CMDB and anchors on CIs
- Presents changes and proposed plans distribution strategy for change and responds to release capacity requests
- Release

Roles and Responsibilities

Storage Capacity focus areas:
- Produce and maintain storage capacity plan
- Monitor storage service levels
- Recommend storage tuning
- Manage storage demand
- Recommend storage capacity enhancements
- Account for workload changes in storage service level targets
Knowledge Statements
Systematic Approach

Assessment is a systematic approach to enable diverse requirements, including availability, capacity, cost, compliance and governance to be met over time. The process achieves alignment at best cost by summarizing requirements and matching them to capabilities at each requirement state change. The range of solutions available are limited by the value of information, with the goal of minimizing budgets, both capital and expense.

\[ f_{(ILM)} = \int_{t_{\text{ingestion}}}^{t_{\text{destruction}}} \lim_{\Delta \text{ information value}} \sum \left( \text{Business Requirements} \rightarrow \text{Storage Capabilities} \right) \]

\[ \text{time to achieve alignment} = 28 \text{ to } 36 \text{ months} \]
Storage Tier

A storage tier is a collection of storage capacity that meets requirements with a consistent set of attributes, capabilities and characteristics which may include:

- Availability
- Performance
- Quality of Service
- Cost

Storage Tier ≠ Storage Service

For example: it is possible that disks in the same array might be treated as different tiers because of RAID Level, data location on a spindle or other characteristics – the choice is yours…
SEC 17a-3, 17a-4 High Performance, Production Storage Service

$23,475.00¹

Meets electronic records storage and retention requirements as specified in SEC 17a-3, the primary rule, and all amendments including 17a-4 under the high performance² requirements of daily floor trading activities with moderate performance in response for retrieval, view and audit requests. Some compliance requirements met include preservation, acceptable media, indelibility, availability, authenticated access, record duplication, chain of custody, index, search and storage at a remote location. This service varies from current amendments by providing a secondary remote storage site to maintain all provisions of 17a-4 during catastrophic events. Also meets Sarbanes/Oxley, Gramm-Leach Bliley and regional privacy requirements

Note #1: Storage Subsystem charges are per TB (1,000,000,000,000 bytes) of allocation per month and are inclusive of all lifecycle management, staff and infrastructure costs.

Note #2: Utilization shall not exceed 50% of allocation to maintain performance profile.
Storage Service
Technical View

SEC 17a-3, 17a-4 High Performance, Production Storage Service

Primary Production Instance

- Start of batch
- End of batch
- Previous day, End of batch
- Next day, Start of batch

- RAID 1/0
- RAID 6 (double parity)
- Fabric Authentication
- Content Aware Storage, Indelible
- Replication, Local
- Replication, Remote
- SNAP or CDP
- CAS Replication
- Link Encryption
- Tape Backups
- Tape Encryption
- Data Destruction

D/R Hot Site

- Start of batch
- End of batch
- Previous day, End of batch
- Next day, Start of batch

Data only Site

- Start of batch
- End of batch
- Previous day, End of batch
- Next day, Start of batch

$23,475.00\text{1}

\text{1} \text{ © 2000-2007 Knowledge Transfer}
Now for something more: *technical* . . .
Basic Storage Capacity

Storage Perspective
- Block
- Storage Pool
  - Given set of capabilities
  - Quality of Service Range
  - Pools are described by:
    - Total Managed Storage
    - Remaining Managed Storage
- Primordial Pool
  - Total System Size (Raw)
- Concrete Pool
- Blocks, Metadata and Reported Capacity

User Perspective (SRM)
- Allocated
  - Static
  - Dynamic
  - Virtual
- Available
  - Mapped
  - Unmapped
- Used, Free, Total, Reserved
Basic I/O Metrics

Arrivals, $A$ \[ \rightarrow \] Transactions in the System, $N$ \[ \rightarrow \] Completions, $C$

Response Time, $R$

Arrival Rate $= \gamma = A \div T$

Throughput $= X = C \div T$

Steady State Throughput $= X = \gamma$
Service Center Metrics
(Queuing Theory)

Service Center, \( i \)

\[
\begin{align*}
\text{Arrivals} & \quad W_i & \quad S_i & \quad \text{Departures} \\
\text{Departures} & \quad R_i
\end{align*}
\]

Graphs:
- Throughput vs. Arrival Rate
- Residence Time vs. Arrival Rate
- Queue Length vs. Arrival Rate
Little’s Law

Response Time Curve

Service Time = 1
Utilization = Service Time + Traffic
Response Time = Service Time / (1 - Utilization)
Little’s Law – Random arrivals and response time

Unit of Work at 0% Channel Utilization

Unit of Work at 50% Utilization

Queue Limitation and impact of Spin Lock

Throughput Capability
- 150 Tops/hour/machine
Financial (Machine)
- $150,000 purchase
- $30,000/year annual maintenance

Throughput Capability
- 4 Batches/hour/machine
Financial (Machine)
- $40,000/purchase
- $20,000/year annual maintenance

Throughput Capability
- 50 Stools/hour/person
Financial (Person)
- $35,000/person/loaded

Assembly
Storage Performance Requirements

Performance requirements

• I/Os per second (average/peak)
  – Bandwidth (average/peak)
  – Jitter (range/mean+std. Deviation)
  – Isochronous

• Access requirements
  – MTFB (initial access time)
  – Locking (concurrent access)
  – File system affinity

• Security requirements
  – Encryption level
  – Integrity
  – Indelibility

• Load characterization
  – Random read vs. write %
  – Sequential read vs. write %
  – Block size distribution

Other requirements

• Retention
  – Retention period/Expiration date
  – Disposition (erase/shred/archive, …)

• Scalability (is this an SLO?)
  – Initial size (+Max. size)
  – Size growth % per time period
  – Bandwidth growth % per time period

• Format
  – Record vs. byte stream vs. ?
  – Media portability

• Co-location/Separation
  (for association)
  – Level (host/datacenter/geographic)
  – Co-operation

• Price
Growth

• Growth (Capacity: Size)
  – Normal: Trend, Horizontal, Vertical (account for each)
  – Burst: Aggregation, Consolidation, Integration
  – Special Cases: Line Items/Transactions

• Performance (Capacity: Performance, Workload Response)
  – General metric is response time (RT)

• Availability

• Manageability (Financial: Cost)
  – How many people does it take?
Defining Scalability

![Graph showing scalability with axes for response time (ms) and throughput (IOPS). The graph illustrates the relationship between response time and throughput, with different load levels and limits.](image)
Q&A / Feedback

• Please send any questions or comments on this presentation to SNIA: trackstoragemgmt@snia.org

Many thanks to the following individuals for their contributions to this tutorial.

SNIA Education Committee

LeRoy Budnik, Knowledge Transfer
Phil Huml, Knowledge Transfer
Bob Rogers, Application Matrix