Northeastern University

Platform Considerations in Fixed Content Clustered Storage Systems

Jim Espy – EMC²

11/14/2006

Outline and Focus

- Overview
- Fixed Content Storage Cluster
  - Components and System Attributes
  - Discussion with Comparisons to Block Storage
- Summary and Take-Away Thoughts

Overview – The Big Picture

Some Top Level Goals and Requirements

- Access to and Protection of Valued Information
  - Store and retrieve information when needed anywhere and anytime
  - Expectation of reliability, availability and security – SLO's and SLA's

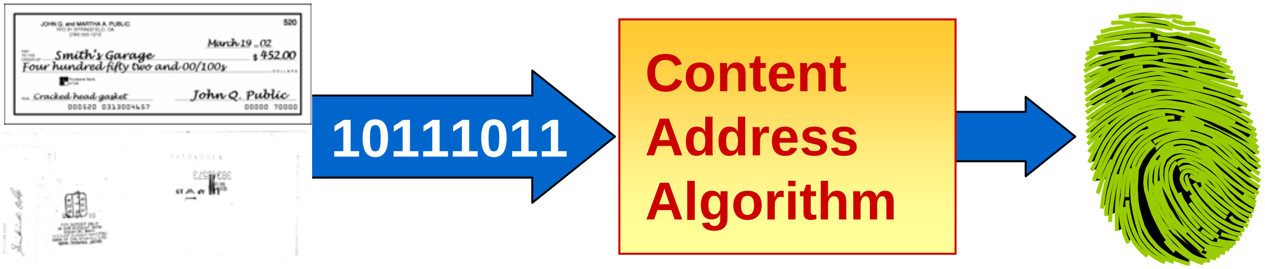
- Managed Information
  - Policy-based access
  - ILM – Tiered storage

- ➢ Specific Need to Store/Archive Reference Data
  - Fixed content
  - Data never changes
  - Requires proof of immutability

nvtech.com

Proposed Solution

- Object Approach to Storage
  - Store data and metadata
  - Utilize hash coding to produce unique naming for stored items
  - Allows for location independence and data redundancy elimination - SIS

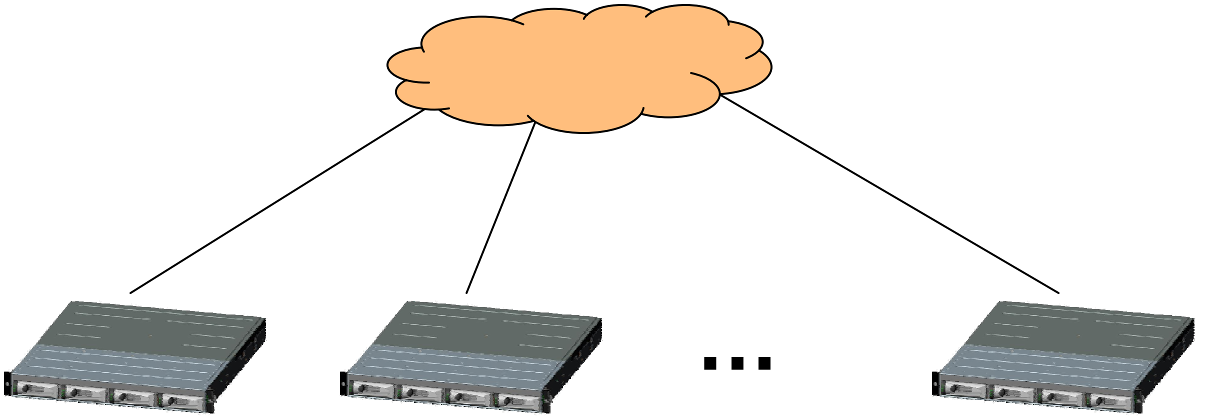


- Cluster Implementation
  - Primarily a software solution – Platform abstraction
  - Interconnected low cost scaleout commodity storage nodes – Purpose-built
  - System-level approach to data reliability, availability and load balancing
  - Self-healing storage without need for immediate service
  - Minimized management
  - Non-disruptive upgrade

Some Key Platform Measurements

- Very Low Purchase and Upgrade Cost
- Very Low Ownership Cost
- Nearline Tier Performance
- Scalability to Small and Very Large Systems
- High Data Availability
- High Data Reliability
- Simplified Serviceability
- High Level of Platform Abstraction

Fixed Content Storage Cluster



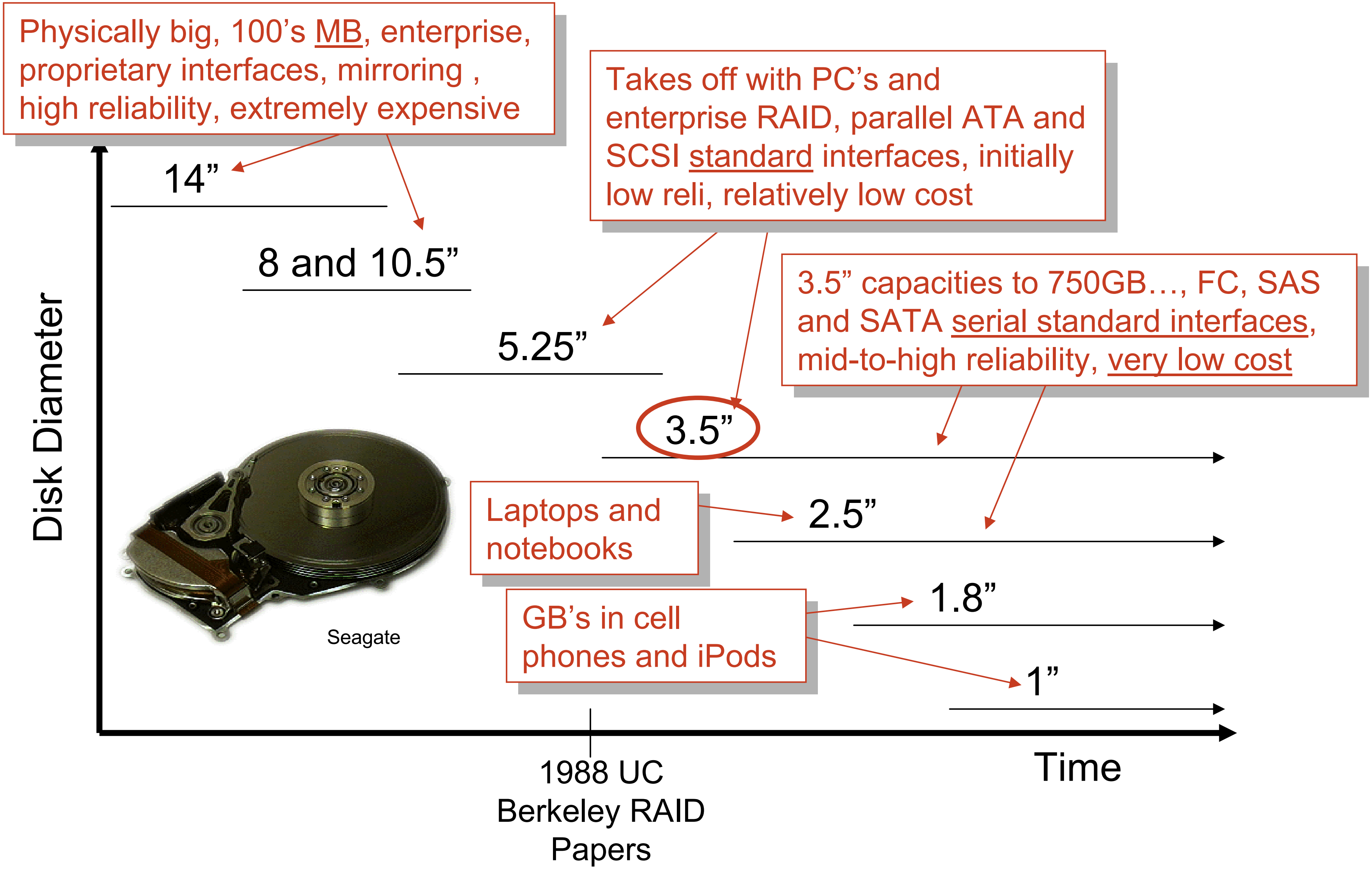
Enablers: Leverage Commoditization

- Just What are These Commodity Components?
  - Driven by high volume commercial markets – Not the enterprise market
- Disk Drives
  - High capacity devices @ 100’s GBs – For just a few hundred bucks
  - Gamers, DVR, iPods, cell phones
  - Flashrom gaining - Also due to iPods and cell phones
- Processors and I/O Devices
  - PC and low-end server-class boxes – Linux and Windows
- Networking
  - Ethernet ports (wired and wireless) nearly as ubiquitous as AC wall outlets
  - Low cost switches and routers
- Can Literally Buy This Stuff Anywhere
  - But what Best Buy sells isn’t exactly enterprise class

Enablers: Leverage Connectivity

- Connection Standardization a Key Enabler
  - Once all these commodity devices could interconnect, every took off
  - Standards and serialization
  - Sub-system chip interconnect to networked storage
- Scalability

Drive Diameters → High Volume & Low Cost



Physically big, 100's MB, enterprise, proprietary interfaces, mirroring , high reliability, extremely expensive

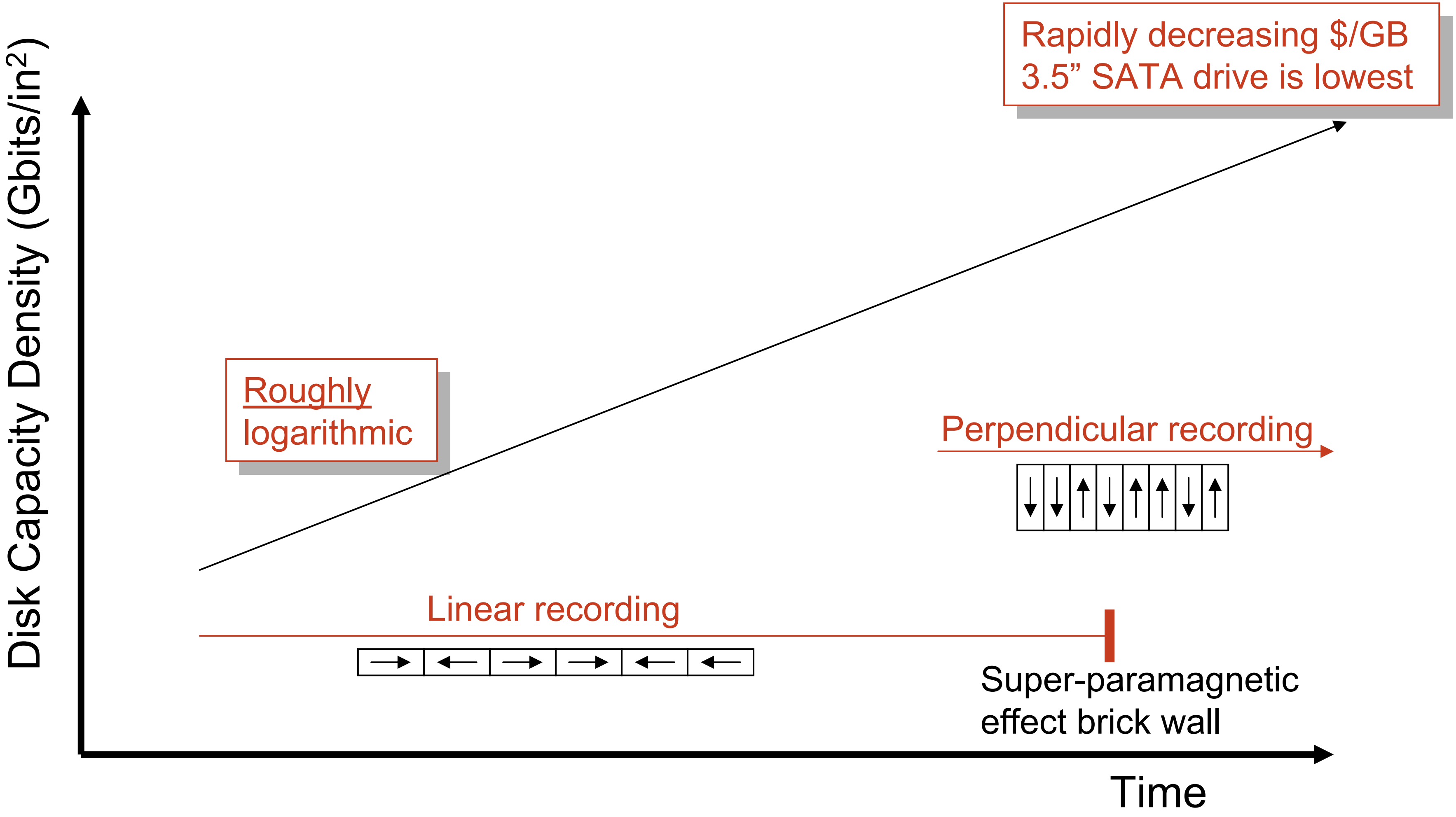
Takes off with PC's and enterprise RAID, parallel ATA and SCSI standard interfaces, initially low reli, relatively low cost

3.5" capacities to 750GB…, FC, SAS and SATA serial standard interfaces, mid-to-high reliability, very low cost

Laptops and notebooks

GB's in cell phones and iPods

Disk Drive Capacities and Recording Technology



Rapidly decreasing $/GB
3.5" SATA drive is lowest

Roughly
logarithmic

Perpendicular recording

Linear recording

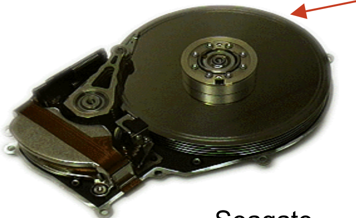
Super-paramagnetic
effect brick wall

Disk Drive Parameters – Changing with Time

On-Line:
Higher cost
Higher reliability
Higher RV resilience

Seagate

Actually 2.5” diameter platters


Seagate

Fewer platters than 3.5”

3.5” SAS and FC dual port, 32MB
Moderate capacity (up to 300GB)
High performance (15KRPM)

2.5” SAS dual port, 16MB
Lower capacity (up to 146GB)
Moderate performance (10KRPM)

Near-Line:
Desktop:
Laptop:
Lower cost
Lower reliability
Lower RV resilience

Seagate

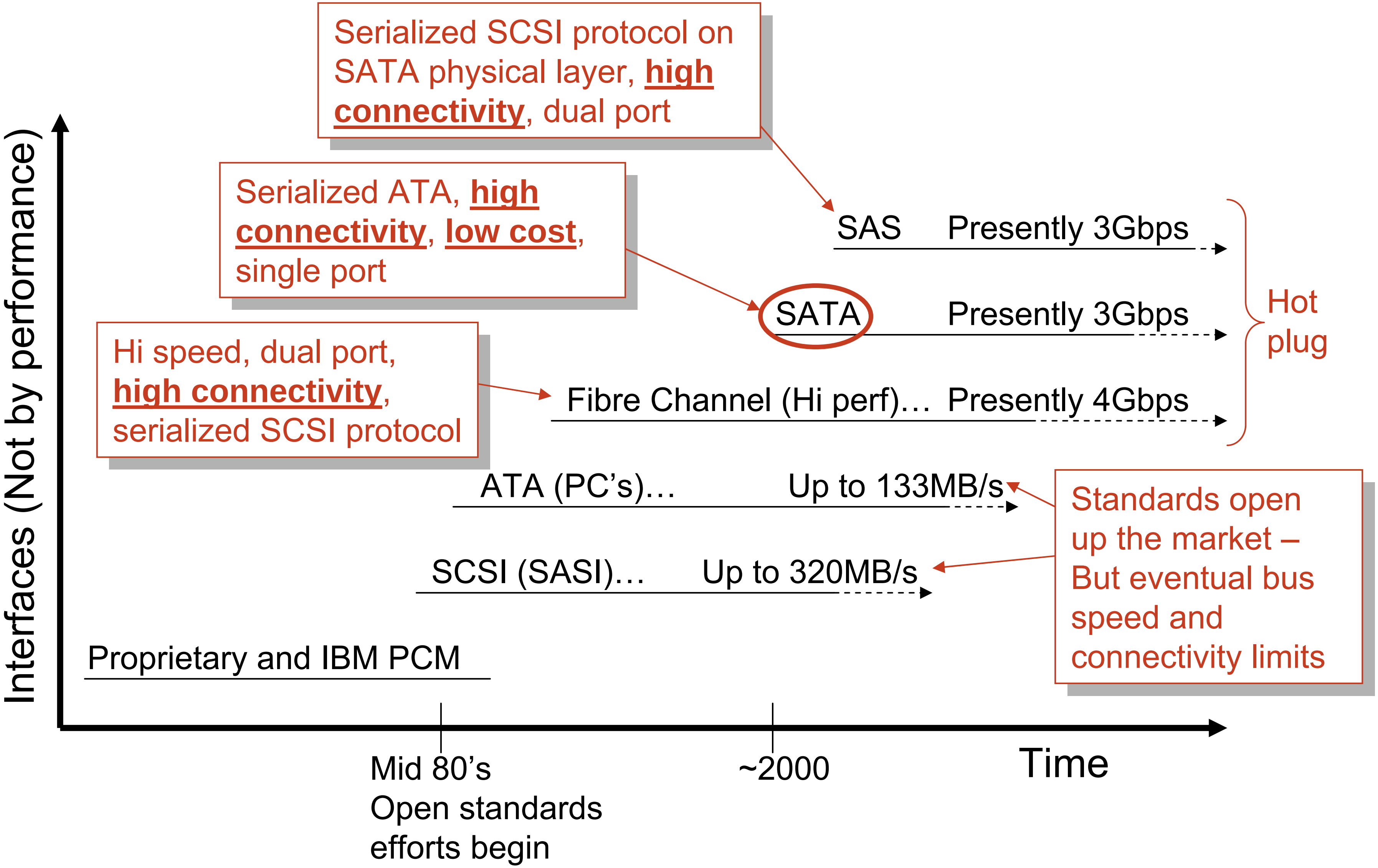
SATA drives dominating sales volumes

Seagate

3.5” SATA single port, 16MB
Very high capacity (up to **750GB**)
Moderate performance (7200RPM)

2.5” SATA single port, 8MB
High capacity (up to 160GB)
Moderate performance (5400RPM)

Disk Drive Interfaces – Towards Standards



Serialized SCSI protocol on SATA physical layer, **high connectivity**, dual port

Serialized ATA, **high connectivity**, **low cost**, single port

Hi speed, dual port, **high connectivity**, serialized SCSI protocol

SAS Presently 3Gbps

SATA Presently 3Gbps

Fibre Channel (Hi perf)… Presently 4Gbps

Hot plug

ATA (PC's)… Up to 133MB/s

SCSI (SASI)… Up to 320MB/s

Standards open up the market – But eventual bus speed and connectivity limits

Proprietary and IBM PCM

Interfaces (Not by performance)

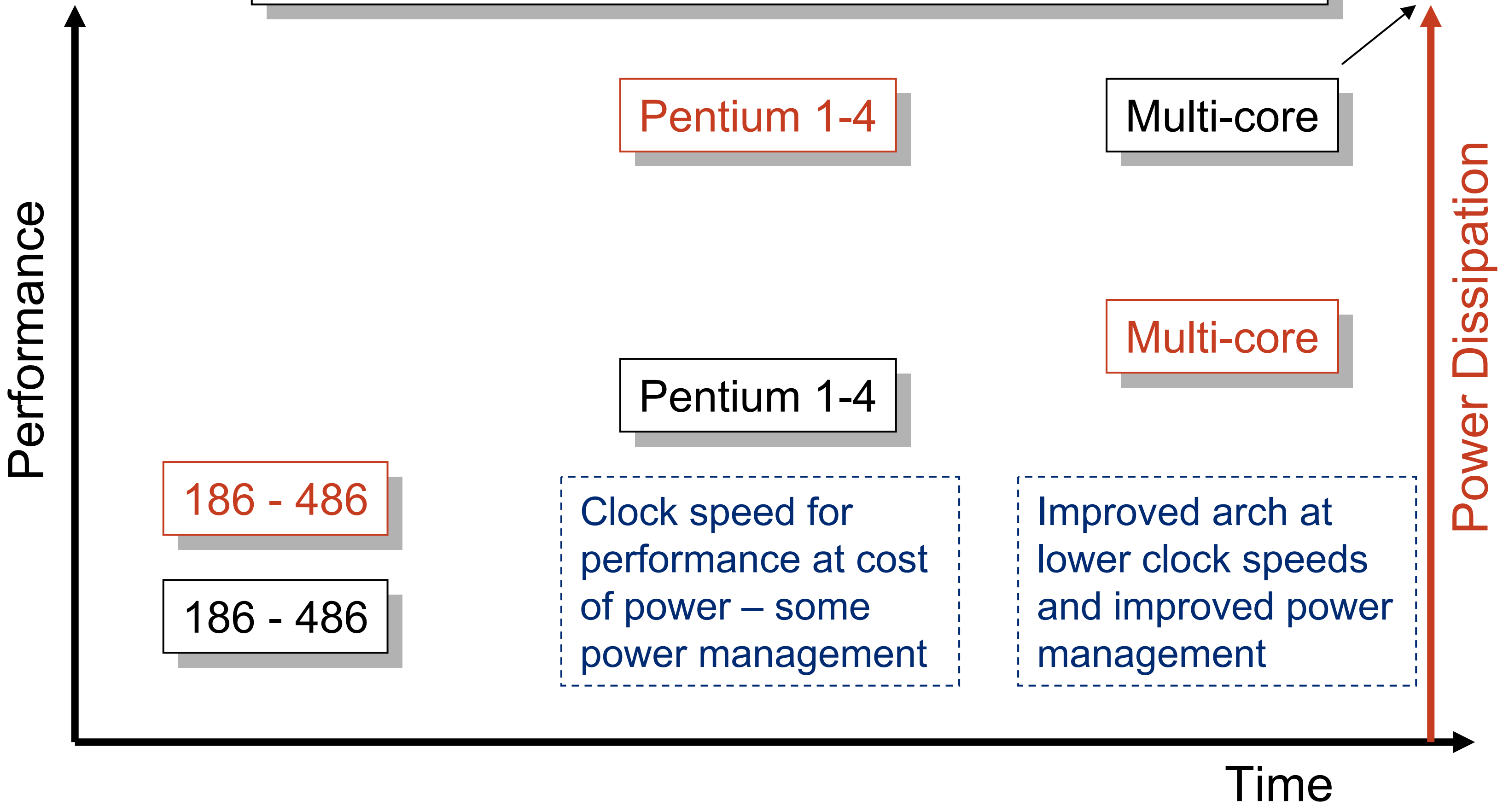
Mid 80's
Open standards
efforts begin

~2000

Time

Processor Trends - Intel

Could also utilize AMD and PwrPC with claims of equal performance at lower power. But who has more commodity platforms and is as entrenched with Linux?



Performance

Pentium 1-4

Multi-core

Multi-core

Pentium 1-4

186 - 486

186 - 486

Clock speed for performance at cost of power – some power management

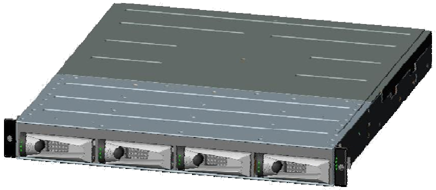
Improved arch at lower clock speeds and improved power management

Power Dissipation

Time

More Processor Trends

- Linux
  - Increasing use in server markets
  - SMP, drive hot plug, etc. support
- Support Chip Architectures and Serialized I/O
  - Faster FSB and memory speeds – ECC support
  - Integrated drive interfaces – ATA initially, now SATA
  - Move from parallel PCI-X to 2.5Gbps PCI-express - Going to 5Gbps
  - Also HyperTransport
- Numerous PC and Server Class Modules
  - Rackmount form factors – Also commoditized
  - Various cost/performance levels
  - Most with multiple drive slots – SAS and SATA



1U = 1.75”

Storage Connectivity Trends

- Moved From Parallel Busses to Serial Connect
- SCSI/Fibre Channel Enables Connectivity & Scaling
  - Creation of Storage Area Networks – SAN's using specialized FC switches
  - 1Gbps → 2Gbps → 4Gbps → Expecting 8-10Gbps
  - But it's considered relatively expensive
- Possible Lower Cost – SCSI, TCP/IP and Ethernet
  - iSCSI SAN's using low cost commodity Ethernet switches
  - 1Gbps → 10Gbps → ??
  - Stack has a lot of overhead – Requires costly off-load engines
- Infiniband Also Vying for Attention
  - Originally hoped to take over - Now "mostly" an HPC cluster connect

➢ But Cluster Interconnect Could Just be TCP/IP
  - Lowest cost of all with lots of routing and aggregation features

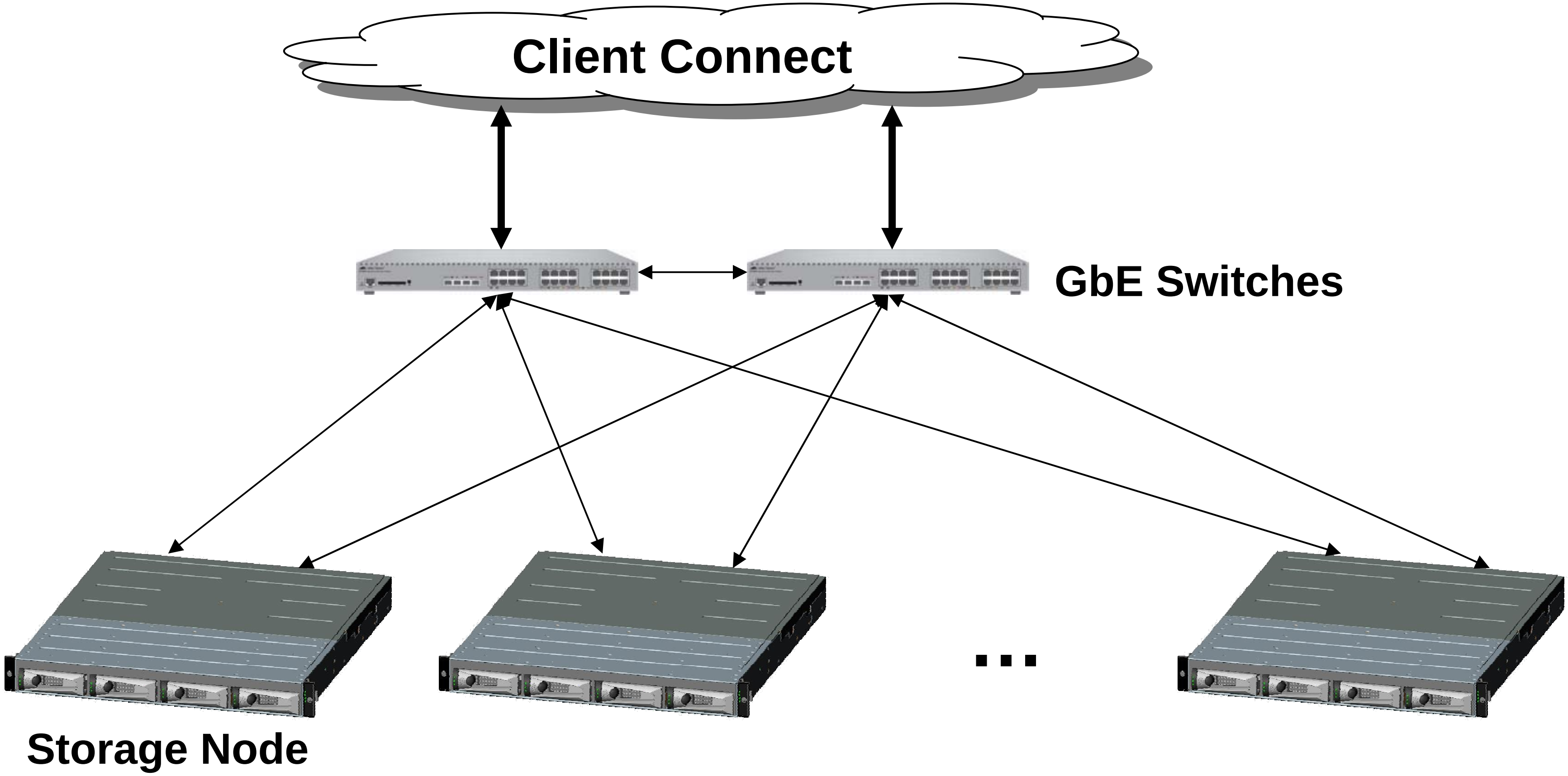
So…

- Why Not Just Network Some Storage PC's Together?
  - Well, that's essentially what Google and others did
  - Low cost and simple
  - What could go wrong?

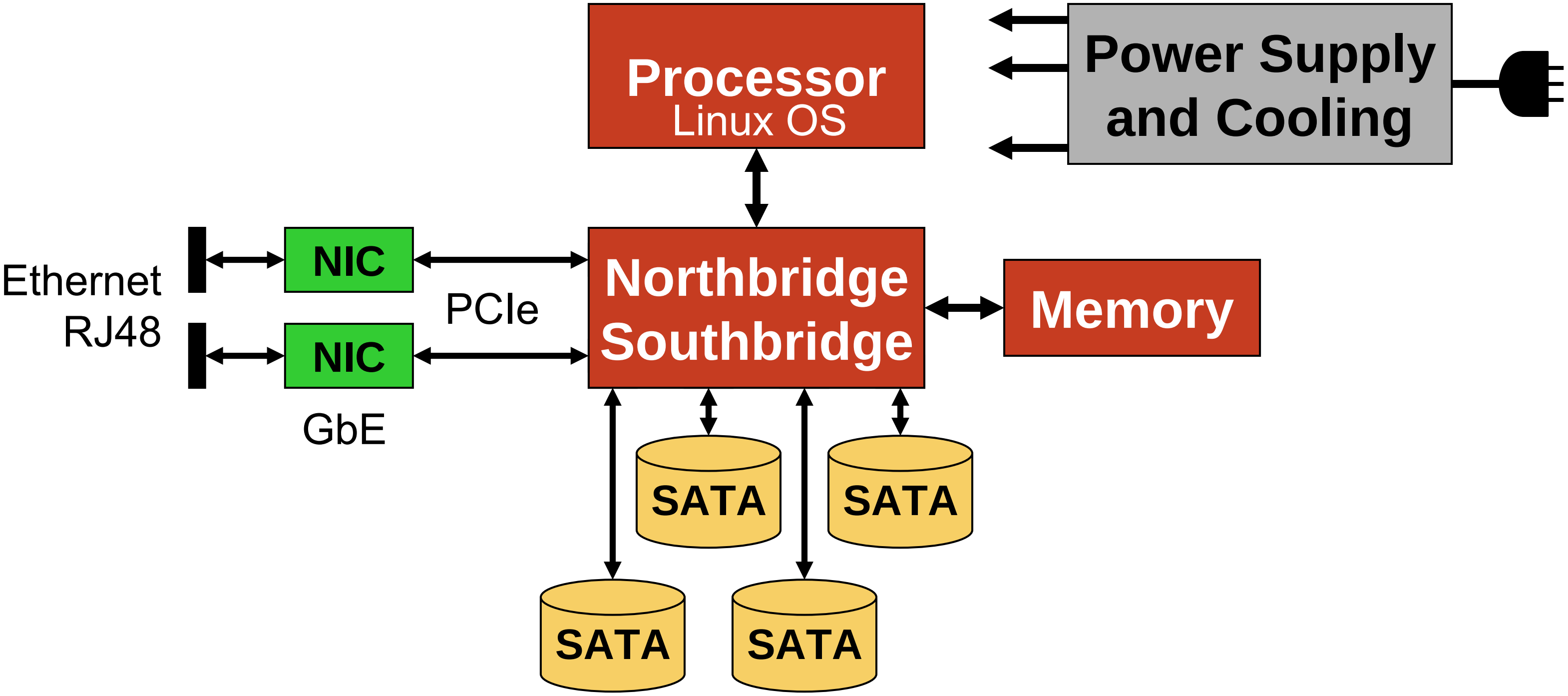
- …Because It's About System Design
  - How to best take advantage of what is essentially a software architecture?

- Some Myths:
  - Disk $/GB is so low that storage is essentially free
  - Processors (PC's) are so inexpensive that processing is essentially free
  - Networks are so ubiquitous that connectivity is essentially free

Scaleout Cluster Storage Interconnect - GbE



Commodity-Based Cluster Storage Node



Commodity-Based Storage Node

- Well Known Configurations
  - Basically a simple non-HA DAS device
  - 500GB drives and growing – Front accessible and cooler air
  - Fairly powerful processors - Around 1GB memory
  - Standardized abstraction mechanisms – IPMI
- But You Get What You Pay For
  - Low cost SATA drives vs. FC or SAS drives
  - Lower reliability, more sensitive, smaller queues and cache
  - Drives originally intended for low duty cycle operation
  - Half duplex and slimmer feature set
- Resource Utilization
  - Take advantage of what is paid for
  - Everyone always want more bloody memory!!

Commodity-Based Network Interconnect

- Keep it Simple
  - Well known tree structures
  - Let software handle the load balancing

- Let the Network do the Heavy Lifting
  - Existing standard protocols – OSPF, LACP, …
  - Ethernet networks designed to handle certain failure situations

- What About Security?
  - ACL’s, NAC, VLAN’s, IPv6, …

System SW-Level Availability

- No Single-Point-of-Failure (SPoF) - So-Called
  - Node, drive, switch or link failure will not result in data unavailability or loss
  - Only explicit cluster HW redundancy involves GbE interconnect
  - <u>Node has no intrinsic HA features</u> – Except dual NIC’s
  - Node failure in a large cluster has low impact on performance

- Expect Failures
  - Effective detection and isolation

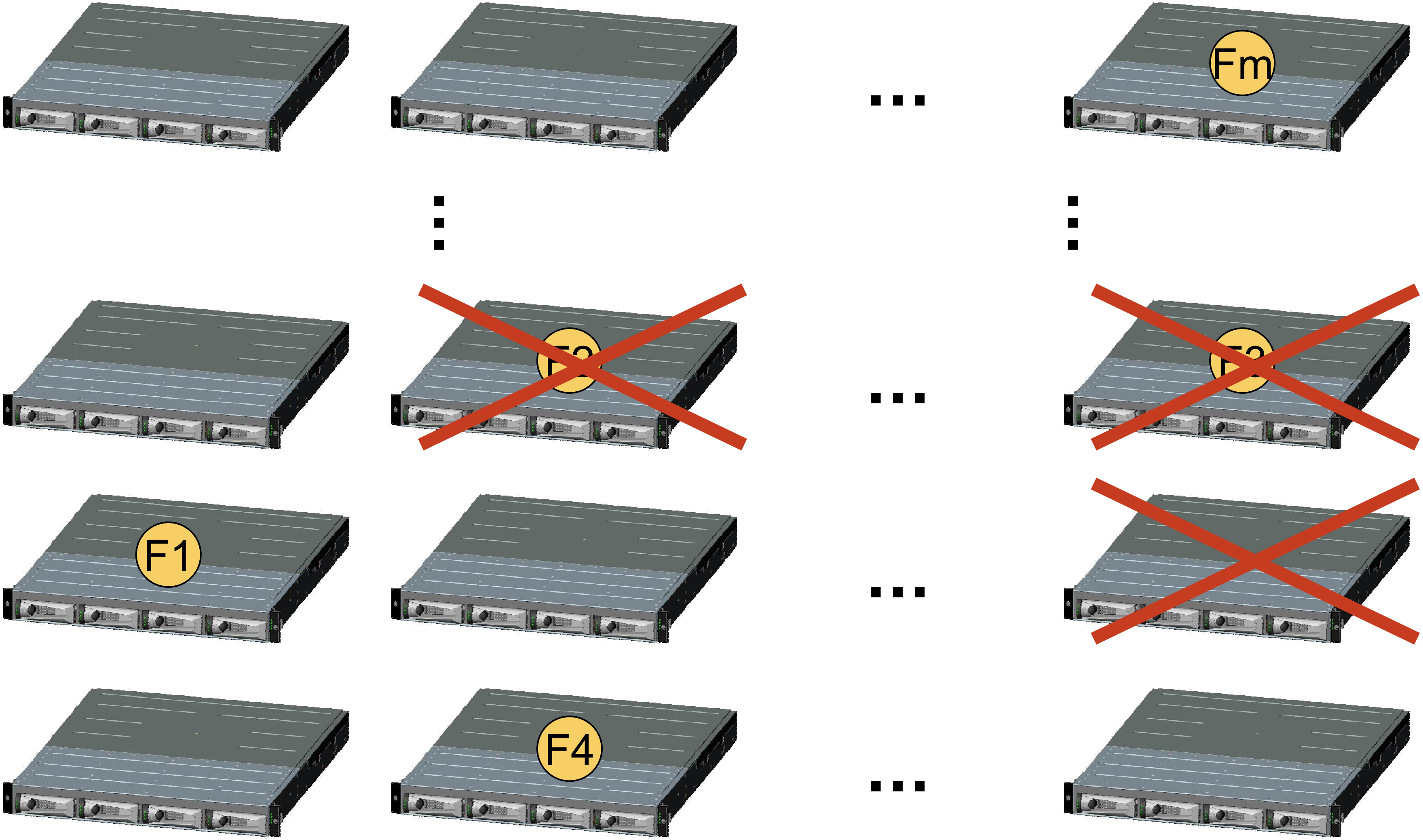
➢ Contrast with Block-Based Dual-SP RAID Arrays

System SW-Level Data Reliability

- RAIN: Redundant Array of Independent Nodes
  - N of M protection where M – N ≥ 1
  - Object to be stored broken into M fragments
  - Redundancy Fragments Distributed Across M Nodes – One per node
  - Any N recovered fragments can reconstruct original object
  - Rebuild is across all nodes utilizing node parallelism for performance
  - Only actual stored fragments need to be rebuilt
  - Sparing is across entire system – No dedicated devices but limited
  - Re-build time vs. “exposure” period
- Also
  - Node busses support parity or ECC – Note that PC memory is not ECC
  - Node serial connections support CRC
  - Hash addressing provides an end-to-end check

➢ Contrast with Block-Based RAID Arrays

Node Failure



Scaling – Performance and Capacity

- Scale to PetaBytes
  - Continue to add nodes – Also enhances performance
  - Issue becomes one of interconnect BW scaling
  - Needs to be non-disruptive
- Also Need to Scale Down
  - Not everyone needs bulk storage
  - Impact on RAIN
- Impact of Increasing Capacity/Spindle

➢ Again, Contrast with Block-Based RAID Arrays

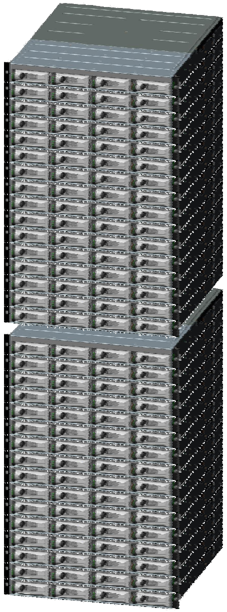
Management and Service

- Lowering the TCO
  - Fewer staff per TB of storage
  - Lower cost service contracts
- Self-Healing Enables Scheduled Service
  - So-called periodic grooming – Scheduled service
  - Enabled by non-dedicated sparing
  - Scheduled service calls cost a lot less than immediate, unplanned ones
- Customer Serviceable?
  - Low number of actual serviceable units
  - Throw-away devices?
  - Let faulty devices die on the vine?

Mechanical Issue - Racking

- Another TCO Issue – Limited Floor Space
  - Maximize storage density

- Stacking in a Rack is Not That Easy
  - SATA drives are more sensitive than FC or SAS drives
  - Sensitive to heat and vibration – Decreasing reliability
  - A big item is Rotational Vibration (RV)
  - Measured in radians/sec^2
  - Stack nodes up and they transmit vibration to each other
  - Possible R/W errors
  - Weight is also a problem – Raised floor limits



Why All the Fuss About Power and Cooling?

- It's Another Big TCO Problem
  - Good news – Fantastic progress in compute and storage densities
  - Bad news – Power densities also increased
  - IT data centers electric bills too high
  - Cooling capacities running up against physical limits

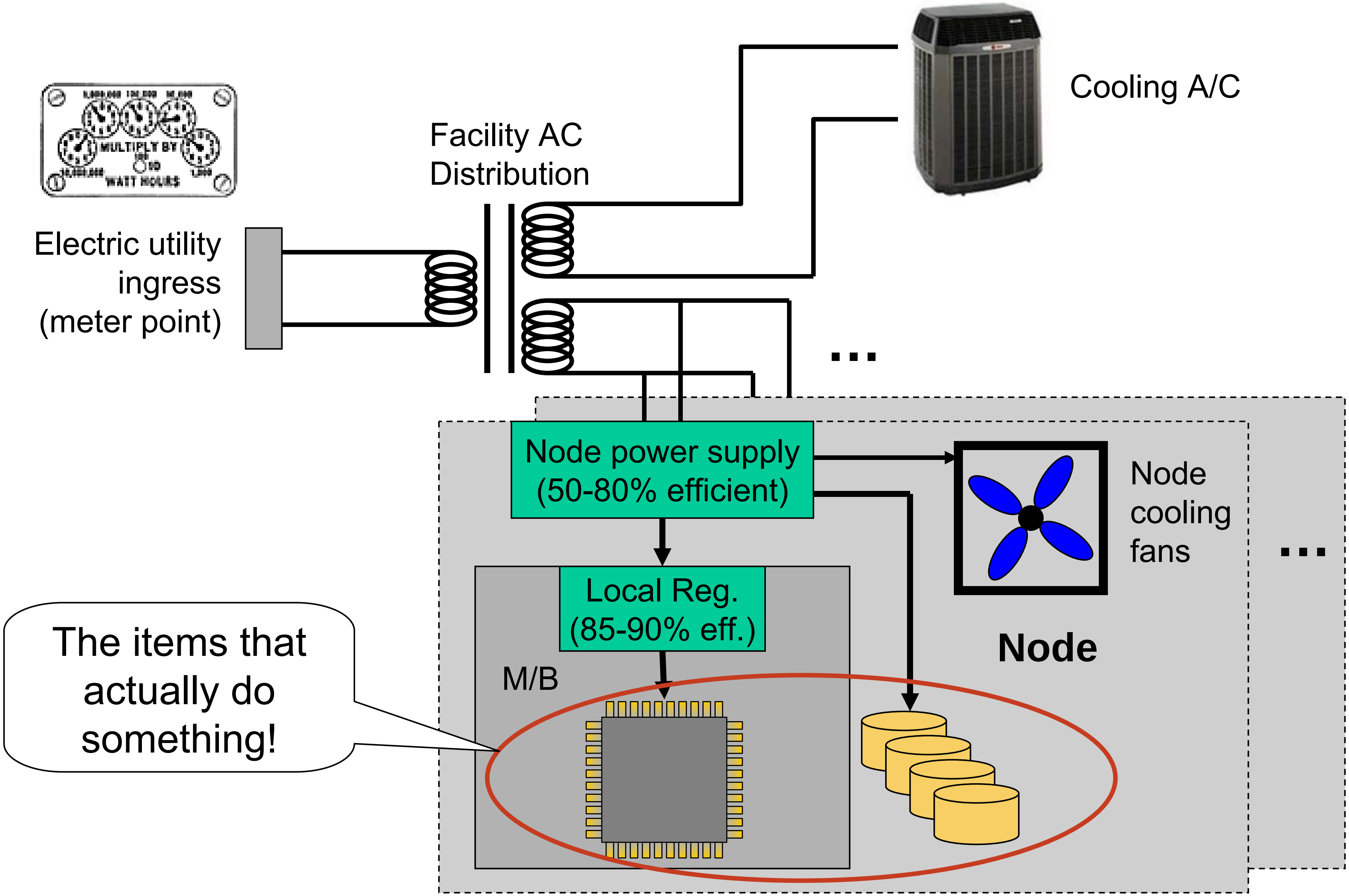
- Made Up Example – Modest ½ Rack Cluster
  - Cost: $75000 - Depreciates over time
  - Power rating: 5000W @ 24/7/365 usage
  - Typical commercial utility rate: $0.14/KWhr (and going up)
  - → 43800KWhrs per year
  - → $6132 electric bill per year

  - But wait…there is ~ 1.5KW data center cooling and overhead per KW used
  - → 109500KWhrs per year
  - → $15330 electric bill per year ← **A decent %/yr of original purchase $**

An Issue of Load and Conversion Efficiency

- Need to Look at the Entire System
  - It’s all about the electric bill and A/C needs/capacity
- Contemporary Server Technology
  - Most devices built for speed, not power savings
  - Unless they were laptop oriented - Solutions expensive
  - Power supplies came from the lowest bidder – Not the most efficient
  - Motherboard standards not optimized for power conversion
- Drives Don’t Help
  - Market push for higher performance - Drive spindle speeds going up
  - A large multiplier
- Data Centers Have Their Own Issues

A Look at a Data Center Power Path



Electric utility ingress (meter point)

Facility AC Distribution

Cooling A/C

...

Node power supply (50-80% efficient)

Local Reg. (85-90% eff.)

M/B

Node cooling fans

Node

...

The items that actually do something!

What Can be Done?

- Lower the Load
  - Drives and processors
- Improve Conversion Efficiencies
  - Spend a bit more and re-architect
  - ATX standard updates – Google proposal
- Improve Distribution
  - DC proposed – Expensive for existing facilities to convert
- Congress Actually Getting Involved
  - Proposing Energy Star* Ratings

What About MAID?

- Massive Array of Inexpensive Disks
  - Potential for very large power savings
  - But only if one has a rather large number of drives to begin with
- Controlled Power Shutdown
  - Drives and/or processing - Predicated on acceptable latencies
  - Wake up still faster than tape latencies - <10 seconds
- Requires Careful Management
  - Reliable re-start of large number of drives - Scary
  - Cannot leave drives off for long time periods without checking
- Existing Low-Level Mechanisms
  - Standard drive spin-down commands
  - Intelligent Platform Management Interface - IPMI

Summary Thoughts

Key Items

- HW Platforms Built From Commodity Devices
- But That Doesn’t Preclude Careful System Design
- System-Level Approach to Data Availability
- System-Level Approach to Data Reliability
- New Approaches to Management and Service
- Power Has Become Non-Trivial

EMC^2

where information lives