Role of Linux in High Performance Computing

vendor perspective

Scott McClellan
CTO/HPTCD

Salishan Conference – April ’03

Perspective

- My Role/Background
  - CTO of the HPTCD division in the new HP
  - Background:
    - pm-HP (1984); commercial computing; OS development (lead architect MPE); telecom(carrier-grade severs & HA architect); joined HPTCD in June/’02
- This presentation represents my personal opinion, not the official position of HP.
- HP is
  - HW vendor (servers
  - Total solution provider
  - Including consulting and services
- Obviously, I can’t/don’t speak for other vendors, but…
  - I suspect other vendors have similar perspectives

Definitions…
(Linux)

Linux: two reasonable/prevalent definitions

- General view:
  - Linux is a “non-proprietary” version of a Unix which is
    - Developed and distributed under an Open Source (GPL) license
    - Ubiquitous, supported on a very large number of HW platforms
- Specific view:
  - “Linux = Red Hat, Red Hat = Linux”

For this talk I will focus primarily on the “general view”, though I will discuss some issues related to the “specific view” as they pertain Linux adoption in the broader HPC market.

Definitions…
(High Performance Computing)

HPC: at least two reasonable definitions

- Narrowly defined:
  - Some folks think of HPC as “supercomputing’
- Broadly defined:
  - Computationally intensive workloads (eg: modeling and simulation, scientific research, etc)
- At HP we take a broader view…
  - Several segments with HPC

In this talk I will focus on the broader definition.

Opinion: taking the broader view greatly benefits true supercomputing customers.

Why? HPTCD is not a HW. We do have influence on the HW roadmaps. The broader HPC market helps us build effective business cases for HPC features.

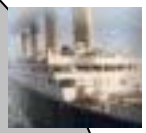
HPC Market
(@ a glance)

Scientific Research

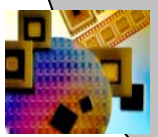
Mechanical Engineering / Virtual Prototyping

Geo Sciences

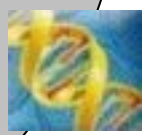
HPC problems are characterized by computational, data-intensive, or numerically intensive tasks involving complex computations with large data sets requiring exceptionally fast throughput.



High-End Film and Video

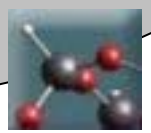


Electronic Design Automation



Life and Materials Sciences

Government Classified and Defense

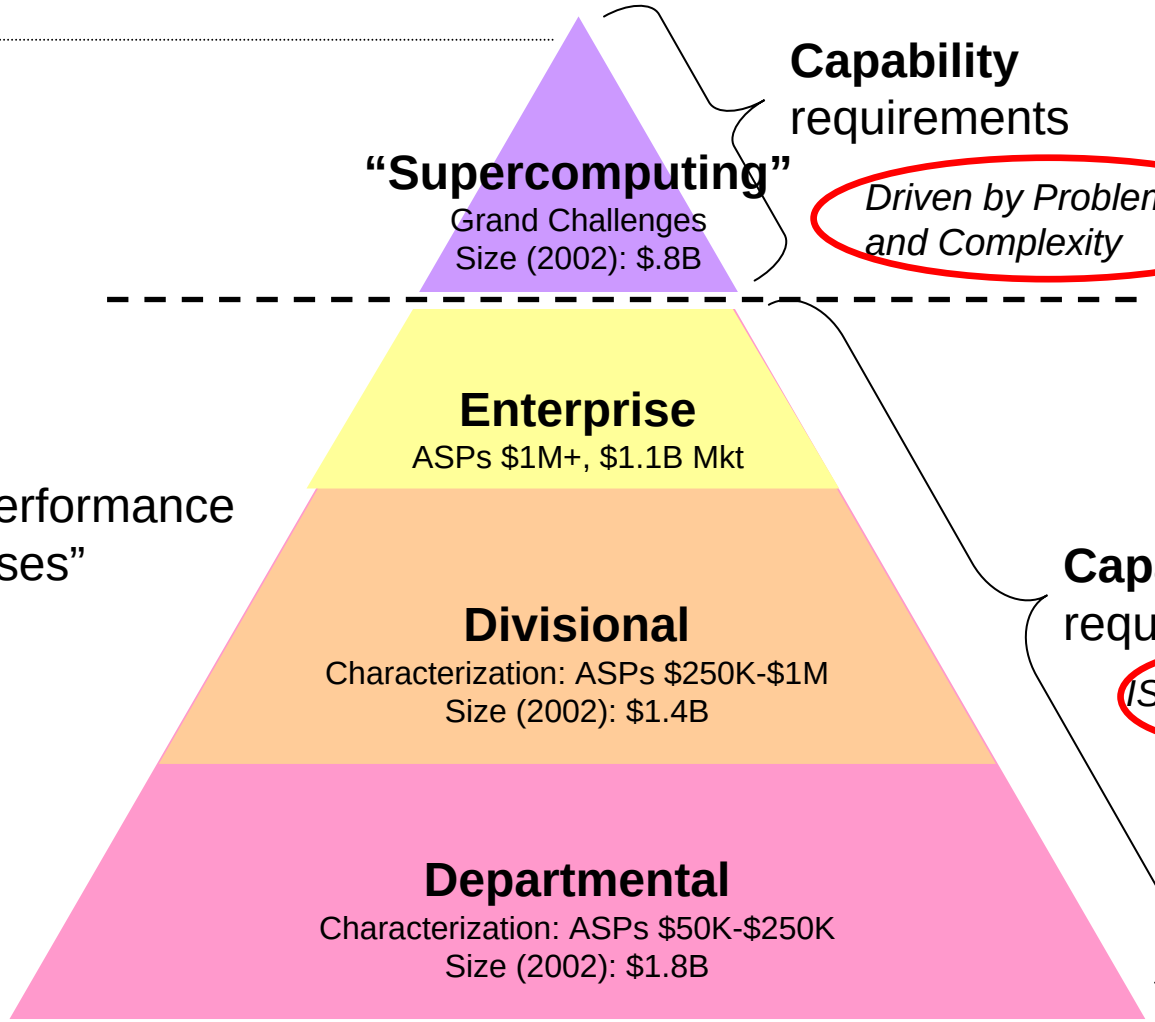


Product Lifecycle Management / Informatics

Performance Class

$5.1B 2001 Technical Computing Market

4 "performance classes"



Capability requirements

Driven by Problem Size and Complexity

Capacity requirements

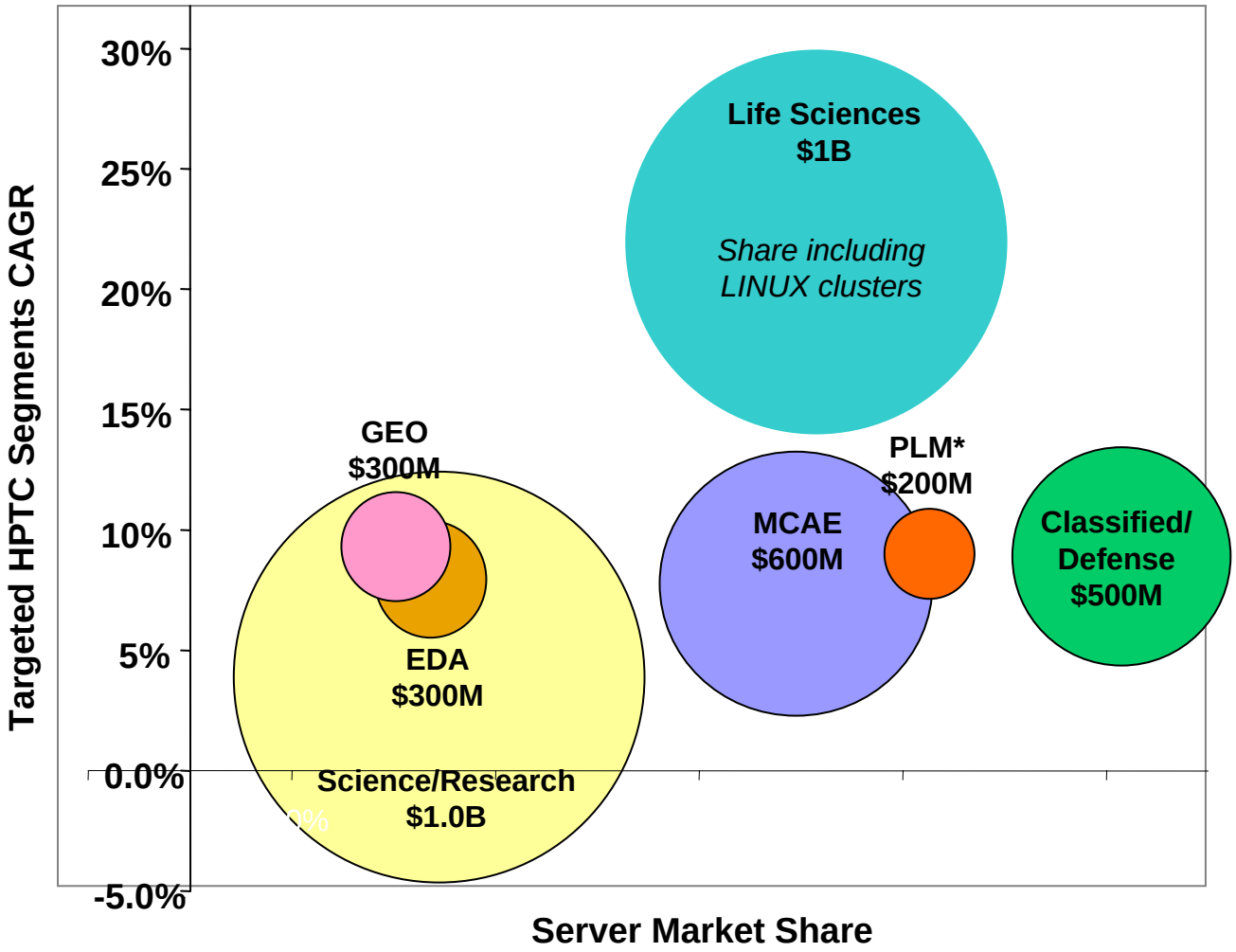
ISV-driven

IDC *Analyze the Future*

Source: August 2002

The factors that drive Linux adoption are quite different above and below the line.

HPTC Targeted Market Segments



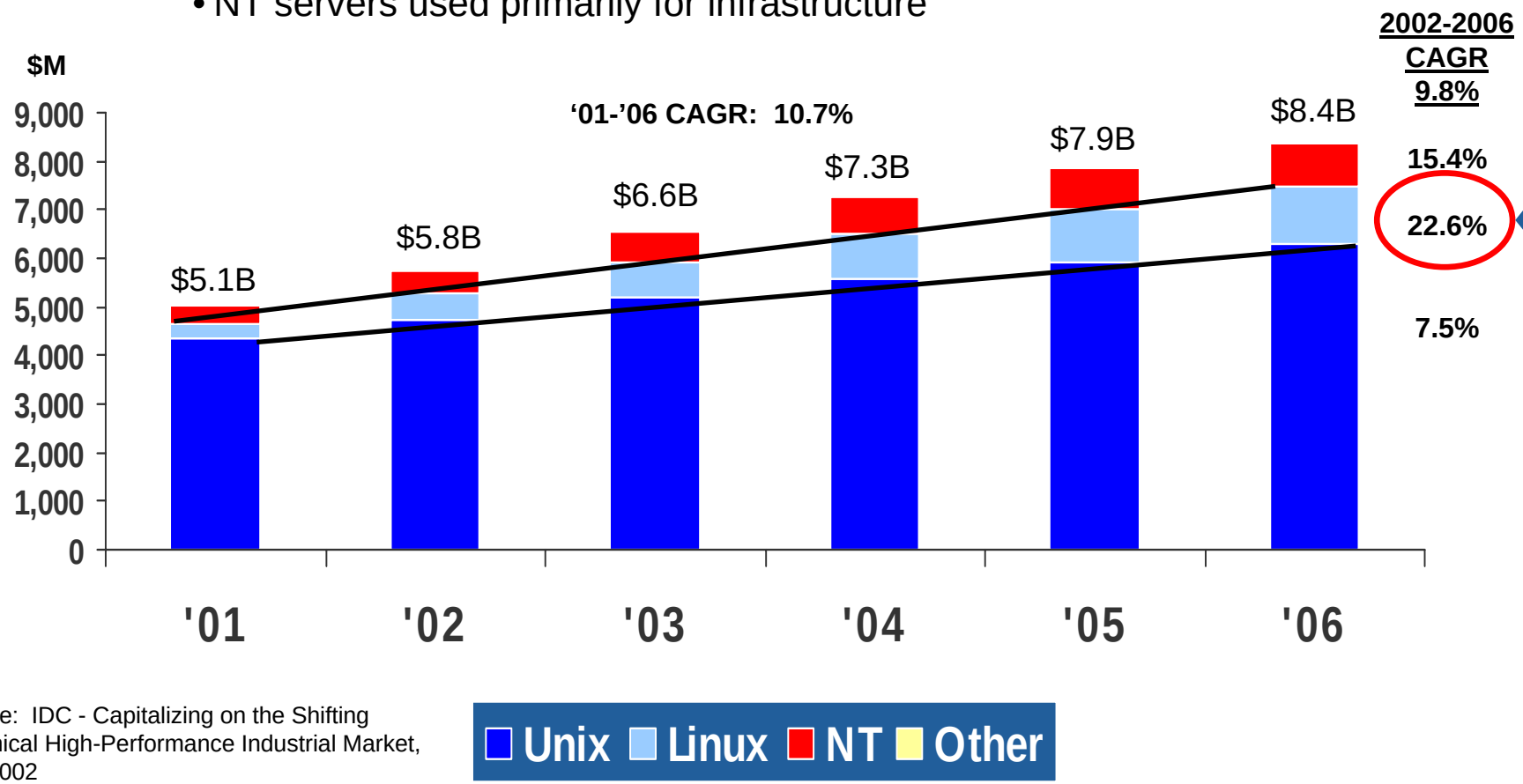
center point of circle indicates CAGR. size of circle represent size of segment.

What is the role of Linux in HPC?

- First the IDC view….

IDC HPTC Market Forecast

- Continued Unix domination
- Emerging interest in Linux as low cost alternative to Unix/RISC
- NT servers used primarily for infrastructure



Source: IDC - Capitalizing on the Shifting Technical High-Performance Industrial Market, Aug 2002

HP believes IDC underestimates the rate of Linux adoption.

What is the role of Linux in HPC?

- Drilling down on the question… from an HP (vendor) perspective

- Key factors influencing Linux adoption
  - Scale out (clustering)
  - Scale up
  - ISV acceptance …
  - Emphasis on “industrial grade’

Linux Enablers
(by market)

Market Segment	Linux Ready	Cluster Ready	ISV Driven	Industrial Grade
Scientific Research	****	*****		
Geo Sciences	****	****		
MCAE	***	****	*****	****
High-end Film	****	****	***	
Life & Material Sciences	****	*****	*****	
Product Lifecycle Management				****
Government Classified & Defense		***		****
Electronic Design Automation			***	****

Clustering
(current state of Linux clustering)

- Several HPC markets have embraced clustering
  - Compelling price performance advantage vs. large SMPs
  - Strong adoption in markets where applications are embarrassingly parallel

- Many Linux clustering “solutions” exist today
  - Completeness and integration varies
  - Mostly middleware solutions
    - Separate, cluster-unaware Linux kernels installed on each node
    - Collection of tools provide the “glue” that holds the cluster together
    - Tools simplify administrative tasks (install, update, configuration, etc); provide monitoring and diagnostics; integrate with MPI, etc
    - Tool vendors have developed parallel debuggers and profiling tools for clusters.
  - Examples: OSCAR, ROCKS, SCORE, + lots of Open Source components
  - Commercial solutions form Scyld, Scali, Linux Networks, MSC Linux

- One problem is there are just too many solutions…

What is going on at the very high end?

- At least four prominent examples:
  - PNNL, LANL, LLNL, SNL
- Observation:
  - All four are very nice designs
  - All four are very different
  - The workloads at the four sites don't appear (to me) to be different enough to warrant four different approaches

- Confusing to a vendor. We would love to jump in an help, but lack of commonality weakens the business case
  - Minimal synergy… minimal opportunity for "re-use"

What Is HP Up To?

- Developing our own Linux clustering solution
  - Called XC, discussed briefly later
  - Focused on providing complete and bullet-proof solution
  - Supplemented with an HP services program
- Developing a state-of-the-art, scalable cluster file system
- Investigating more advanced clustering solution
  - Particularly interested in scalable SSI cluster technology
  - Looking to leverage our technology and industry leading expertise (TruCluster, NSK, VMS, OpenSSI, etc)
  - Technology partnership with Unlimited Scale, Inc.
  - Collaborative efforts with key leading edge customers to maximize technology leverage

XC Program

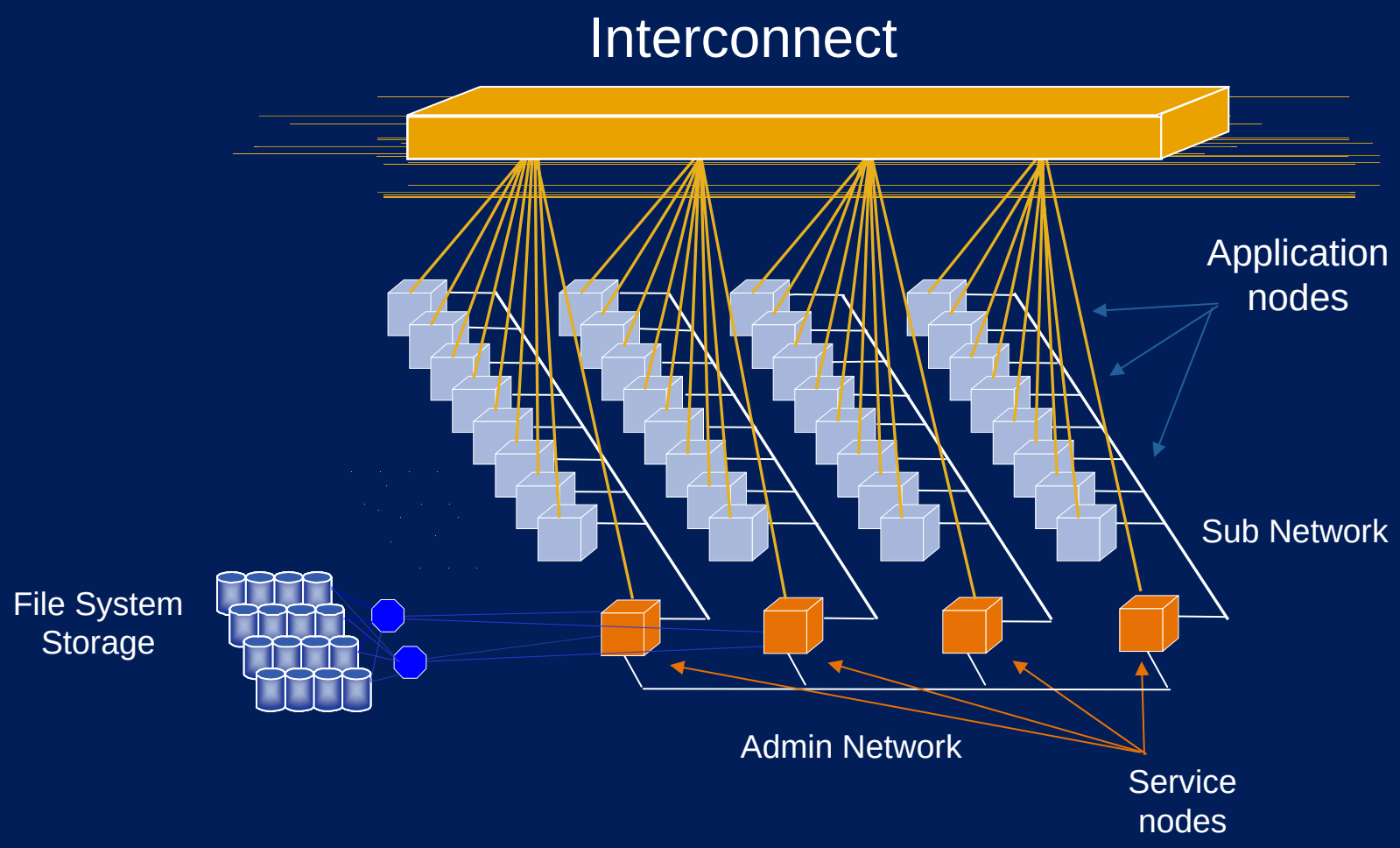
- Deliver a family of industry leading solutions
  - select best of breed technology
  - leverage performance and price-performance advances in core components
  - enable structured flexibility
- Meet the needs of the most demanding technical customers
- Provide ‘industrial-grade’ Linux-based technical solution for the commercial technical user and ISVs
- Production quality product focus
  - Extensive Test cycles
  - Use Field Test to expose system to customer workloads
  - Constrain variations to allow fully tested products

What's all the fuss about?

- A lot of work to "productize" something

- Not a complete list:
  - Create world wide support infrastructure
  - Train support and services organizations
  - Create modular, integrated HW building blocks
  - Deal with HW integration issues like
    - Thermals
    - Cable management; securing cables; observing minimum bend radius, etc
    - Regulatory approval for rack configurations
  - Deal with supply chain issues
  - Create product number sku's to make system orderable
  - [re-] engineer component for modularity, and HW (architecture, interconnect, etc) independence.
  - Determine reference configurations
  - Develop extensive, repeatable certification tests
  - TEST!

High Level XC System Architecture
(XC V1 Product)



XC System Software

- Includes HP customized Linux kernel and Red Hat compatible user environment
- Integrated with key technologies from partners (Unlimited Scale, Inc., Platform, Interconnect drivers)

- HP qualified
- HP documented
- HP supported

Initial platform offerings

- HP xc6000 (name under review)
  - Factory integrated, turnkey cluster of
    - Itanium2-based rx2600 servers
    - Quadrics Elan3
    - XC System software
- HP xc3000 (name under review)
  - Factory integrated, turnkey cluster of
    - ProLiant servers with Xeon processors
    - Myrinet 2000
    - XC System software
- Support and service options

Scale Up

- Ongoing work (multi-vendor cooperation) for Linux to scale up to larger SMP systems.

- Examples:
  - Support for larger page sizes
  - Transparent super-page support
    - Support for very large variable sized pages
  - Improvements in scheduler
  - Reducing lock contention

- Challenges: How to enable ‘proprietary’ features that require kernel level support
  - Example: soft and/or hard partitioning mechanism and other RAS features

Characterizing HP ISV Partnerships

- Application providers in specific markets
  - Including 104 Life Sciences, 68 EDA and 53 CAE application vendors
- Other ISV partnerships
  - Provide technology that completes our solution portfolio
  - System SW and HW vendors
  - SW development tools
  - etc…

- A total 279 partners (and growing)
  - 66 designated as “primary”

Linux Enablers
(by market)

Market Segment	Linux Ready	Cluster Ready	ISV Driven	Industrial Grade
Scientific Research	****	*****		
Geo Sciences	****	****		
MCAE	***	****	*****	****
High-end Film	****	****	***	
Life & Material Sciences	****	*****	*****	
Product Lifecycle Management				****
Government Classified & Defense		***		****
Electronic Design Automation			***	****

Linux Adoption/ISV Challenges
(first mover disadvantage)

- Today vendors work closely with ISVs to
  - Certify, optimize, enhance, support ISV solutions
- Excellent example: CAE
  - Nearly 100% ISV driven market
  - HP works closely with all CAE ISVs
    - Invest resources and $$$ on certification
    - Feed changes back into HPUX and compiler groups
    - Carefully optimize key libraries and ISV code sections
    - Etc.
- First mover disadvantage
  - Takes a considerable effort to support ISV aps on Linux as they are support on HPUX
  - If we tune and certify <xyz>-ap on Linux will we see an ROI? Or will the actual sales go to the cheapest whitebox vendor?

Linux Adoption/ISV Challenges
(Linux = Redhat, Redhat = Linux)

- ISV perspective
  - "platform" = <specific hw config, specific sw stack>
  - Have to "certify" application for each "platform"
  - The specific OS version (sometimes down to the patch level) is a key variable in the equation
  - The will certify on a particular Linux and would (in general) require re-certification if the Linux variable changes
  - Many ISVs have effectively equated the following;
    - Linux = Redhat, Redhat = Linux
    - Implication is that solutions need to be Redhat based (branded) to get broad ISV support.

- Most interesting HPC [clustering] solutions require some kernel modification – even if minor.
  - Additional drivers
  - Kernel hooks for feature xyz (example hooks for cluster file system)
  - Generally speaking, if you modify it at all, Redhat says it is not Redhat anymore.

Some Observations
(and a plug for LSB compliance)

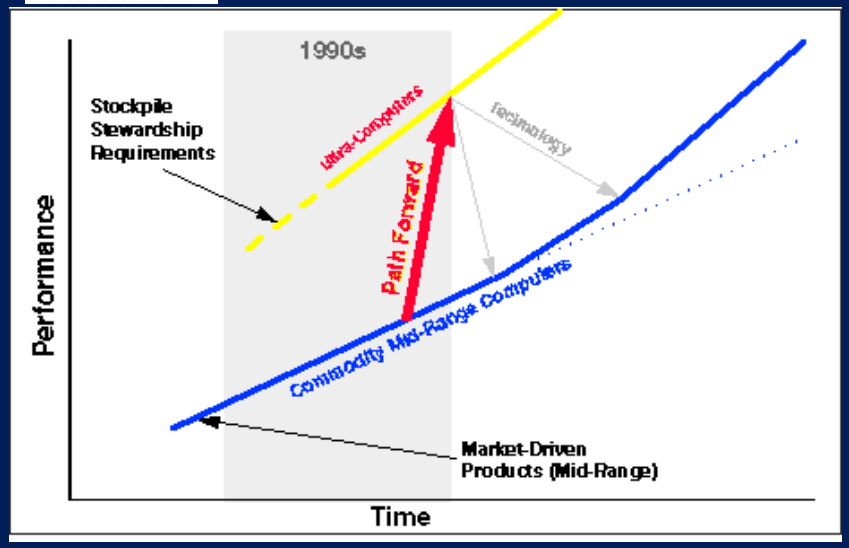
- Linux = Red Hat, Red Hat = Linux
  - Is an odd side effect considering the Linux movement was propelled by an “anti-Microsoft”
  - Pretty constraining for vendors trying to build open source products/solutions with ISV appeal
- Linux Standards Base
  - From my point of view, LSB compliance would be a better way to insure compatibility
  - And a better “logo” for ISVs to look for
- The current model/mindset works fine for the crowd that wants to ship shrink wrapped Linux with every box.

Lustre Overview

HP Hendrix Project

Lustre Selected by ASCI Pathforward

- HP Prime contractor (w/CFS and Intel)
- HP Storage Division (NSS) Lead
- HPTCD involvement rapidly increasing
- 3yr project
- Committed to delivering Lustre in five phases.



Momentum behind Lustre…

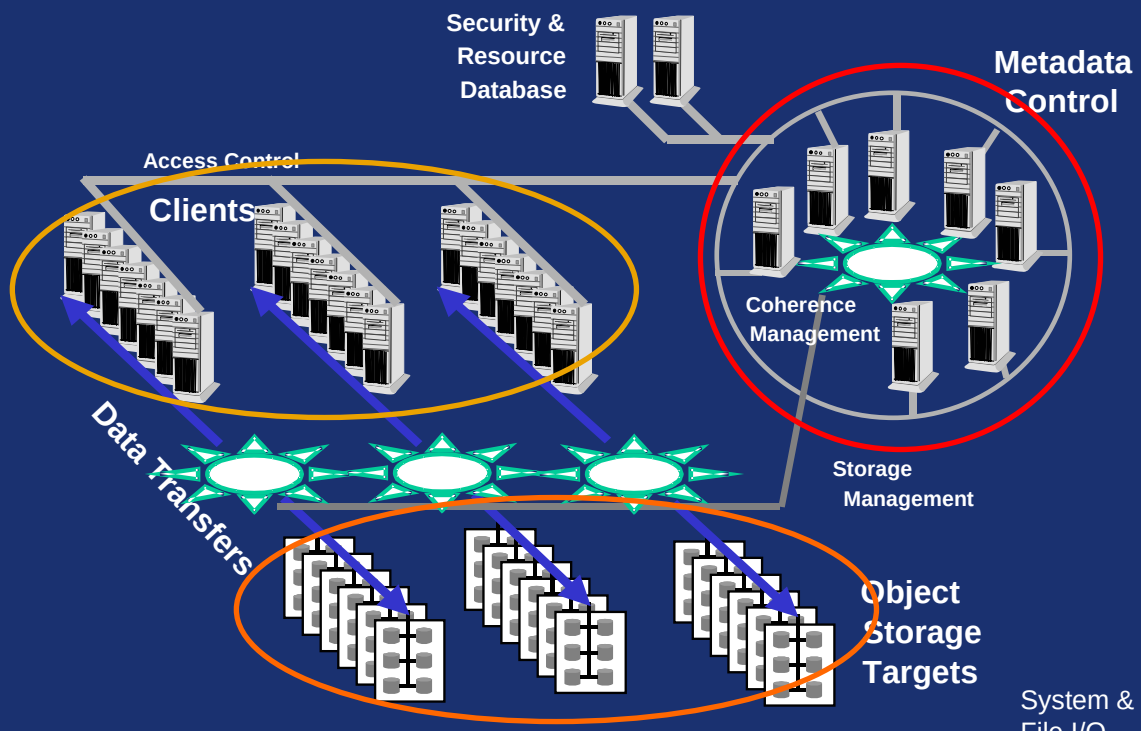
- HP Bid Lustre in ASCI Purple
- HP Hendrix Project – making solid progress!
- Open source - strong interest
- HP committed to Lustre @ PNNL
- LLNL MCR cluster
- DOE & other HPTC customers are excited about a unified file system
- Synergistic w/HP storage division strategy
- Other Storage vendors looking at Lustre!

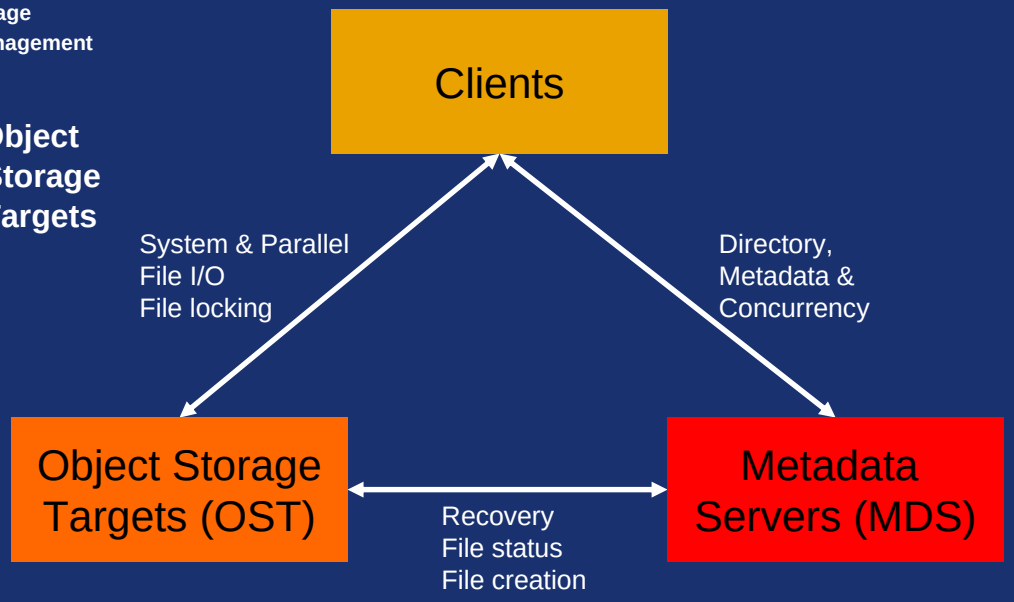
Could become a ubiquitous file system!

Lustre

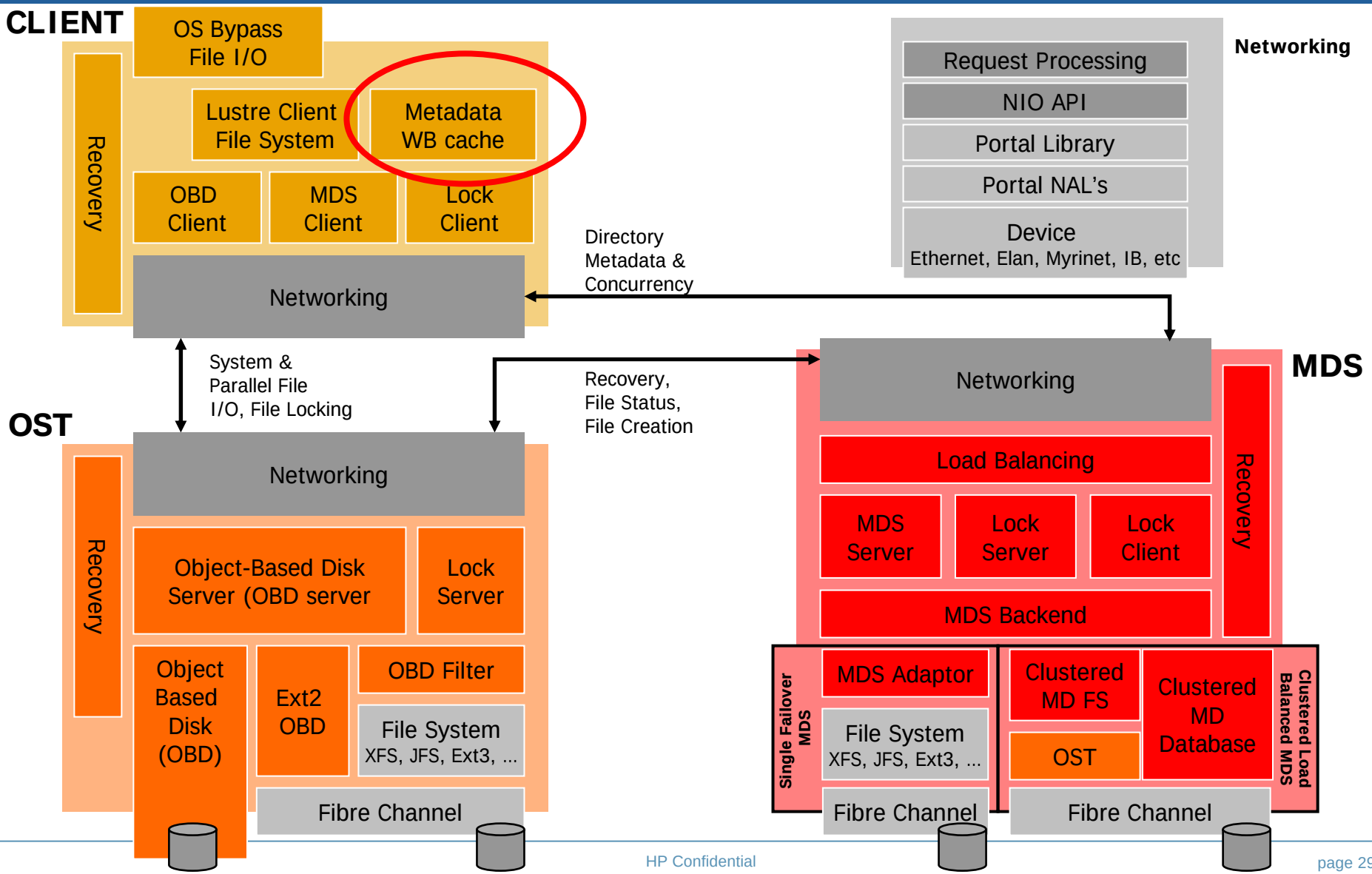
- Two areas of focus at HP
  - HPTCD is looking to Lustre to be *the* highly scalable parallel file system for its scale-out systems
  - HP's Storage Division expects Lustre to be a key part of its NAS offerings
- What's it to you?
  - A way to separate procurements of storage and computes.
  - Scalable!
- What is Lustre's ultimate potential?
  - Ubiquitous (multi-OS, multi-platform, multi-vendor)
  - Heterogeneous

Lustre File System
(logical structure)





Lustre Architecture



Lustre Phased Development
(per ASCI PathForward/Hendrix agreement)

New
0.5

2003

Phase 1 | Phase 2 | Phase 3 | Phase 4 | Phase 5

Lustre Lite

Key Features:

- Functional file system
- Parallel I/O, Distributed benchmarks can be run.
- **100 clients, 10 OSTs, 1 Meta-Data server**
- Performance at 20% theoretical (single node) 10% theoretical (parallel nodes)
- Recover from system failure
- Multiple clients, single MDS
- MDS serving more than 1 file system
- Client node failure recovery automatic.
- Metadata journaling.

LL Performance

Key Features:

- Improve performance to 80% theoretical (single node) 40% theoretical (parallel nodes) on large clusters
- **700 clients, 30 OSTs, 1+1 Meta-Data**
- Scalability is much improved
- Locking system improved.
- Exporting NFS
- MPI-IO support.
- OST fail over

Lustre T10 Security

Key Features:

- Add T10/OSD compatibility and Security pieces.
- Improve Parallel I/O performance @ 70% theoretical
- **3000 clients, 1000 OSTs, 16 node Meta-Data server**

Clustered MDS

Key Features:

- First implementation of scalable clustered metadata servers
- **3000 clients, 200 OSTs, 4 node MetaData server**
- Improve Parallel I/O performance @ 60% theoretical
- (single node still @ 80%)

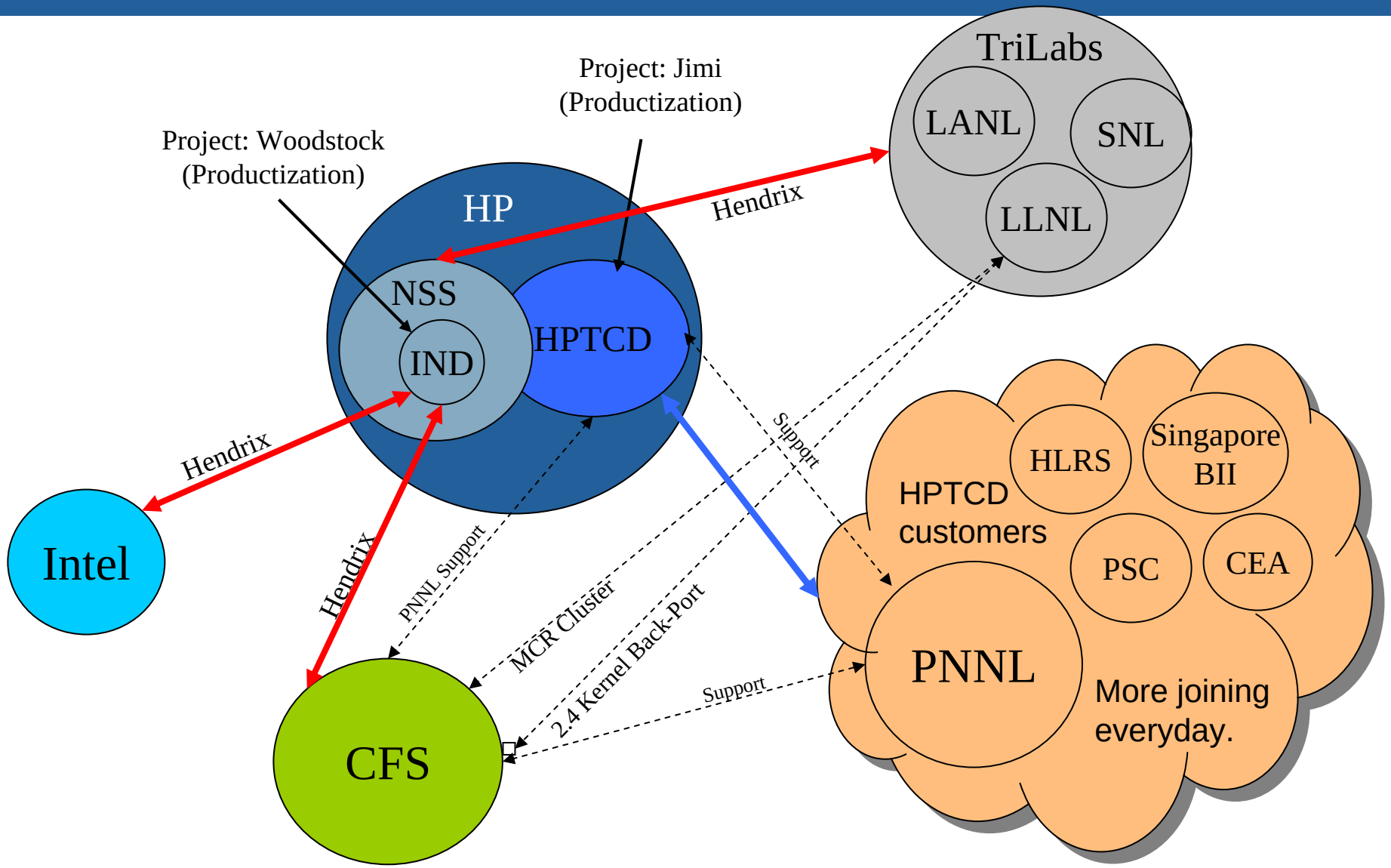
Lustre GNS & Management

Key Features:

- Add global namespace support
- Single node performance to 95% theoretical
- Improve Parallel I/O performance @ 85% theoretical
- **10,000 clients, 10,000 OSTs, 100 node Meta-Data server**
- Integration with enterprise management tools
- Hooks for HSM application

Lustre Lite

The [Rapidly] Growing Lustre Ecosystem



Storage companies…BlueArc

Lustre Overview

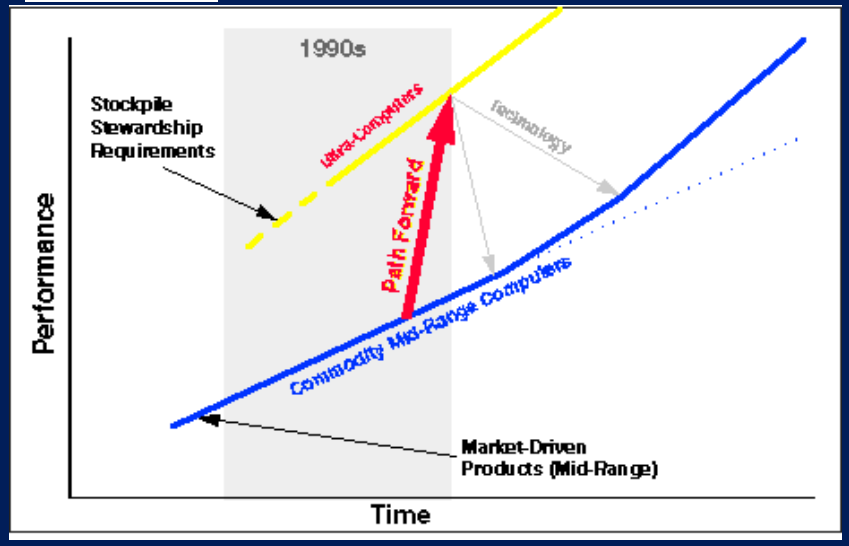
New
0.5

HP Hendrix Project

ASCI Pathforward

Lustre Selected by <u>ASCI Pathforward</u>

- HP Prime contractor (w/CFS and Intel)
- HP Storage Division (NSS) Lead
- HPTCD involvement rapidly increasing
- 3yr project
- Committed to delivering Lustre in five phases.

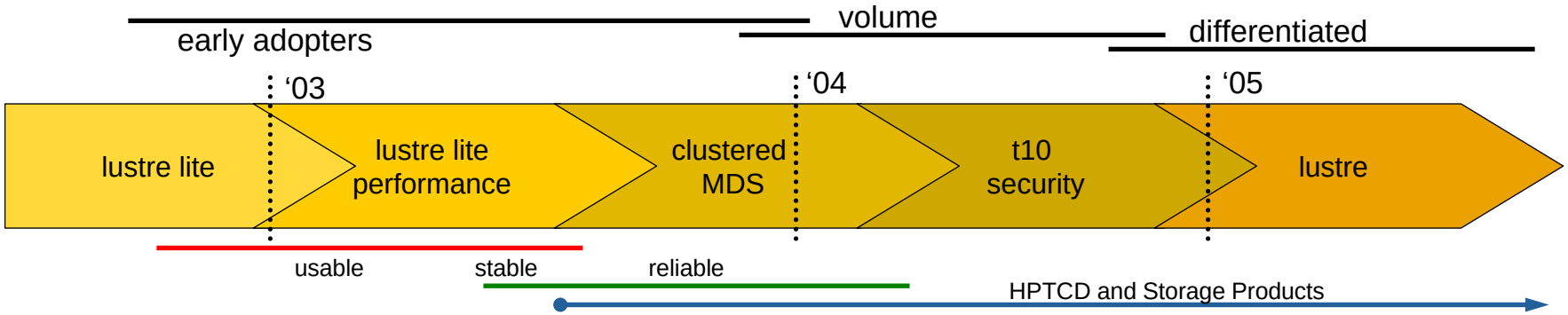


Momentum behind Lustre…

- Pre-HP Bid Lustre in ASCI Purple
- HP Hendrix Project – making solid progress!
- Committed to PNNL
- Customers want it!
- DOE wants unified file system
- LLNL MCR cluster
- Open source, strong interest
- Synergistic w/HP storage division strategy
- Other Storage vendors looking at Lustre
- IBM…
- Microsoft….

Could become a "standard"!

Lustre Roadmap
(setting expectations)



- Lustre will come out in a series of releases
  - Already in use at Lawrence Livermore National Labs
- Early Adopters in 2003
  - Usable
  - Stable
  - Reliable
  - 100x Clients, 10s OSTs, 1+ MDS
- "Volume" in 2004
  - Scalability, ease of management, etc.
  - x000 Clients, x00 OSTs, 4 MDS
- Increased functionality releases into 2005
  - "Product Complete"
  - x0,000 Clients, x0,000 OSTs, 16 MDS

Open Source …
(the good side)

- Very empowering concept…
  - Mechanism for tapping into the collective intelligence of a vast army of developers
  - Linux its is an Open Source success story
  - Lustre is an example

Open Source …
(some challenges)

- Complicates/changes the fundamental business model
  - intellectual property ‘issues’
    - How does one protect any IP
    - IP contamination concerns if mixing open and non-open technology
  - How to make money
    - What can you charge for? How much can you charge?
    - If you have to charge less and gross margins are reduced, then how much can you afford to invest
    - How do you engender durable customer loyalty
- You give up “total control” which effects
  - Lose control of roadmap once technology is open
  - Harder to make specific customer commitments
  - Risks related to depth of internal vs external expertise
    - Complicates support model

Floating an Idea
(…two+ ideas actually)

- Standard for HPC Linux
  - Telco analogy:
- Architecture branch for [tightly coupled] clustering
  - Generally speaking Linux kernel maintainers and Linus are not interested in clustering.
  - Very difficult to get clustering enhancements into standard Linux if they are not needed for the desktop. (Possible but an uphill battle)
- The notion that Linux SHOULD fracture is a defensible notion.
  - Embedded, enterprise, desktop, etc have VERY different design centers.
  - Difficult to impossible for one OS to serve all masters faithfully/optimally.