NFS: What’s Next

David L. Black, Ph.D.

Senior Technologist

EMC Corporation

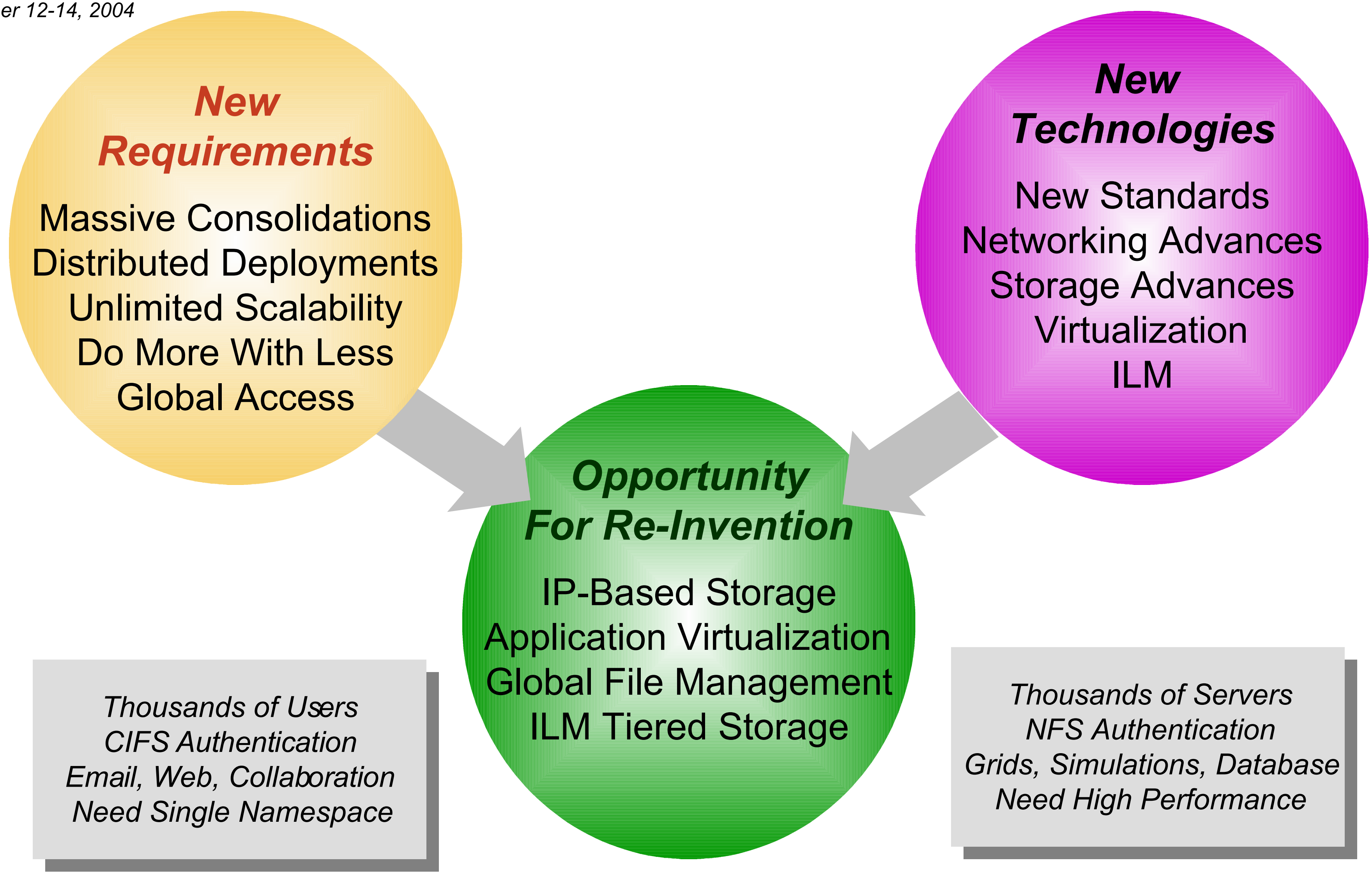
black_david@emc.com

We Briefly Interrupt this Presentation ...

EMC makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

... We Now Return to the Regularly Scheduled Presentation

Networked Storage is Changing….



New Requirements

Massive Consolidations
Distributed Deployments
Unlimited Scalability
Do More With Less
Global Access

New Technologies

New Standards
Networking Advances
Storage Advances
Virtualization
ILM

Opportunity For Re-Invention

IP-Based Storage
Application Virtualization
Global File Management
ILM Tiered Storage

Thousands of Users
CIFS Authentication
Email, Web, Collaboration
Need Single Namespace

Thousands of Servers
NFS Authentication
Grids, Simulations, Database
Need High Performance

EMC Network Attached Storage [NAS] Vision

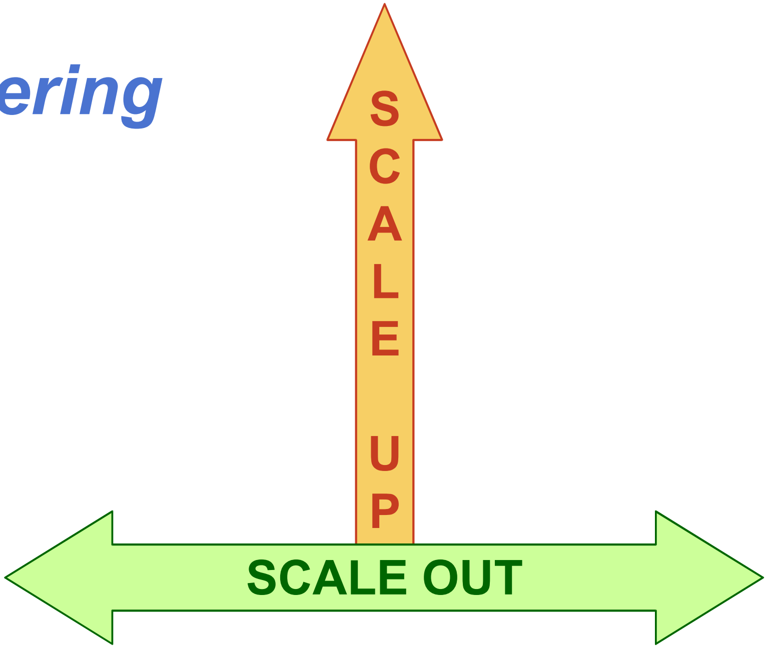
IP-Based Storage Delivering

- Infinite Scalability

EMC NAS Vision

IP-Based Storage Delivering

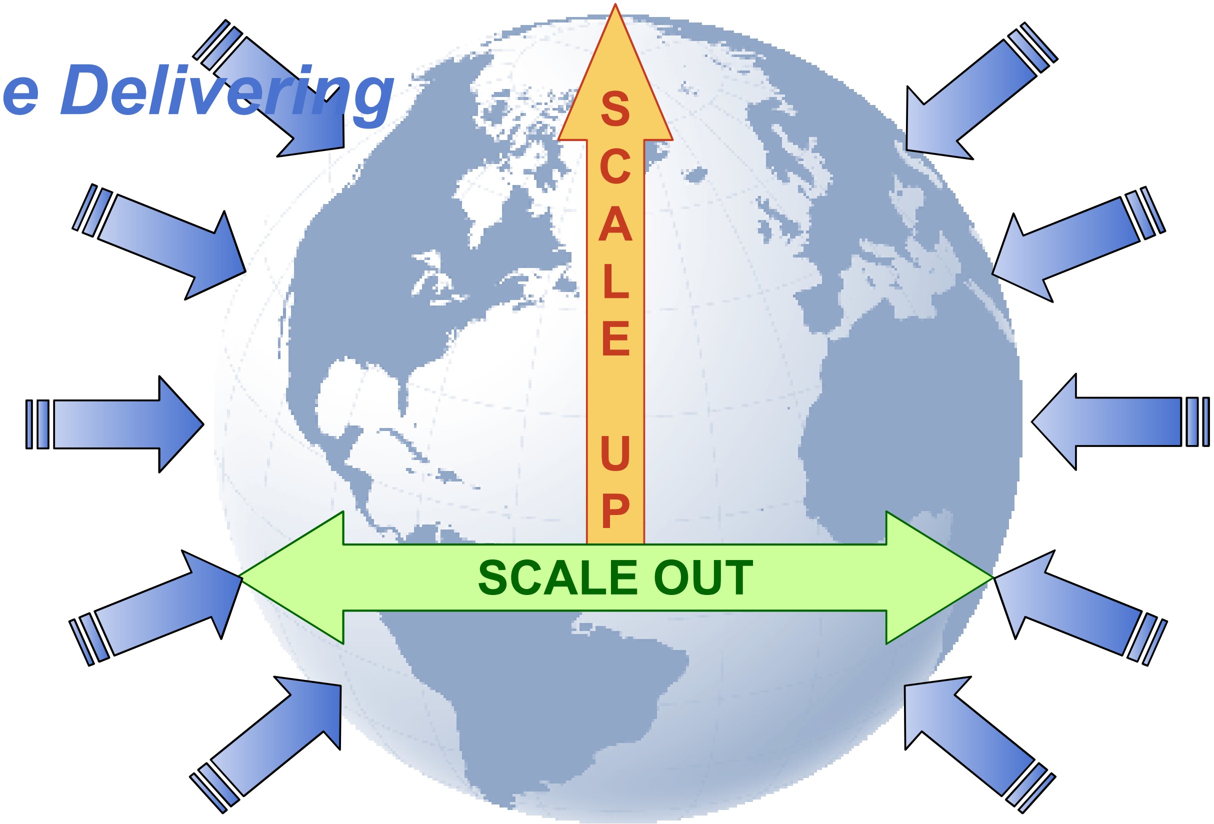
- Infinite Scalability
- Optimized Data Placement



EMC NAS Vision

IP-Based Storage Delivering

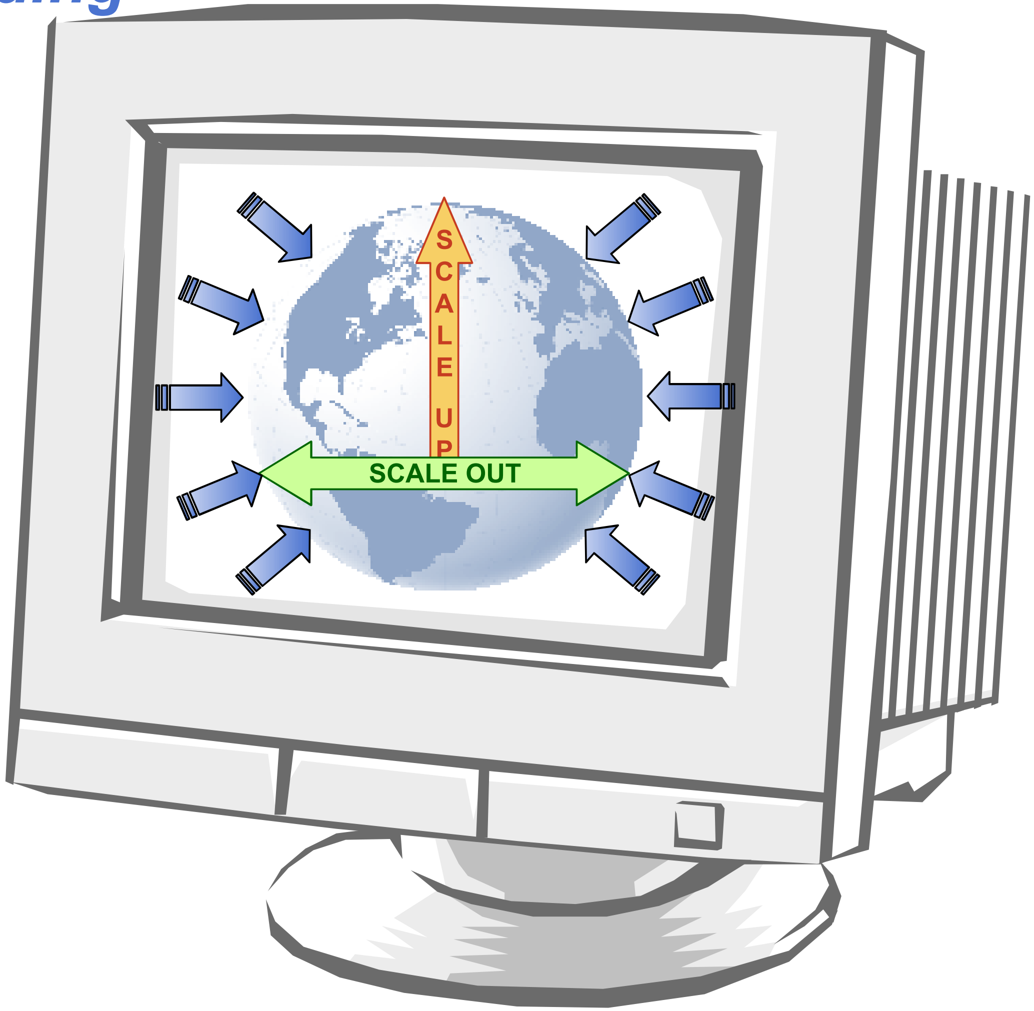
- Infinite Scalability
- Optimized Data Placement
- Global Accessibility



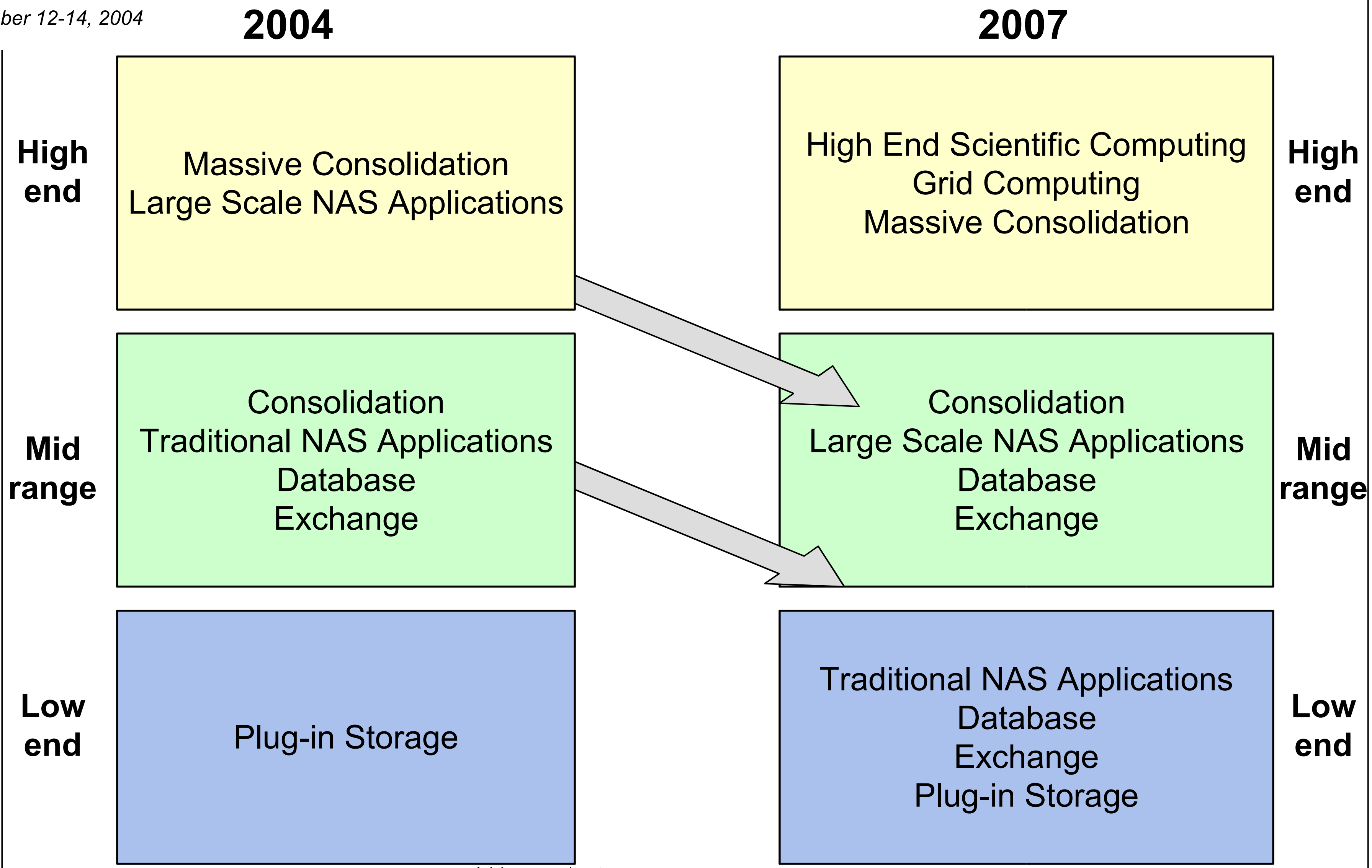
EMC NAS Vision

IP-Based Storage Providing

- Infinite Scalability
- Optimized Data Placement
- Global Accessibility
- Centralized Management



NAS Use and Applications



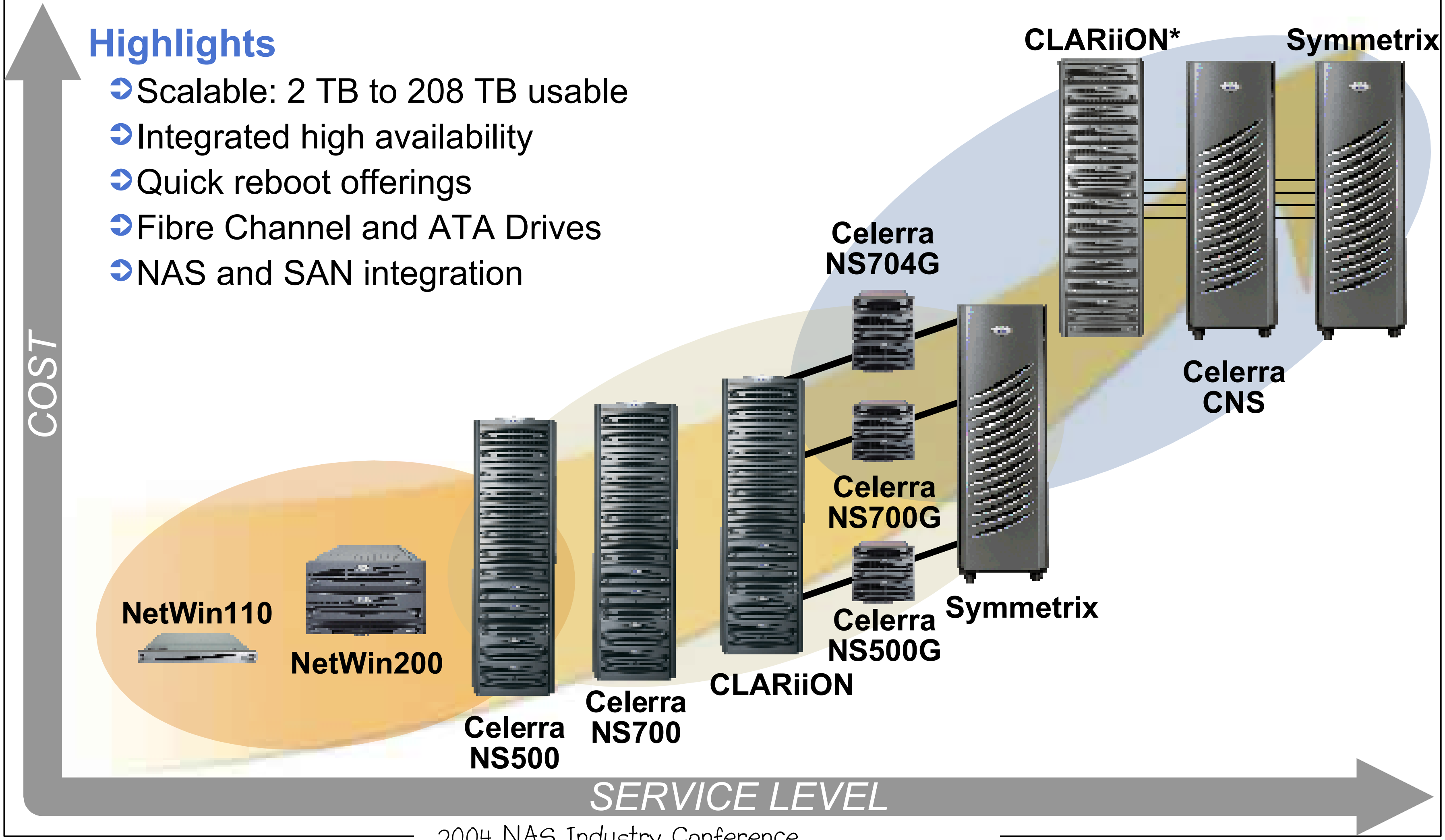
	2004	2007	
High end	Massive Consolidation Large Scale NAS Applications	High End Scientific Computing Grid Computing Massive Consolidation	High end
Mid range	Consolidation Traditional NAS Applications Database Exchange	Consolidation Large Scale NAS Applications Database Exchange	Mid range
Low end	Plug-in Storage	Traditional NAS Applications Database Exchange Plug-in Storage	Low end

EMC NAS Family

Tiered Storage to Meet Different Service-Level Requirements

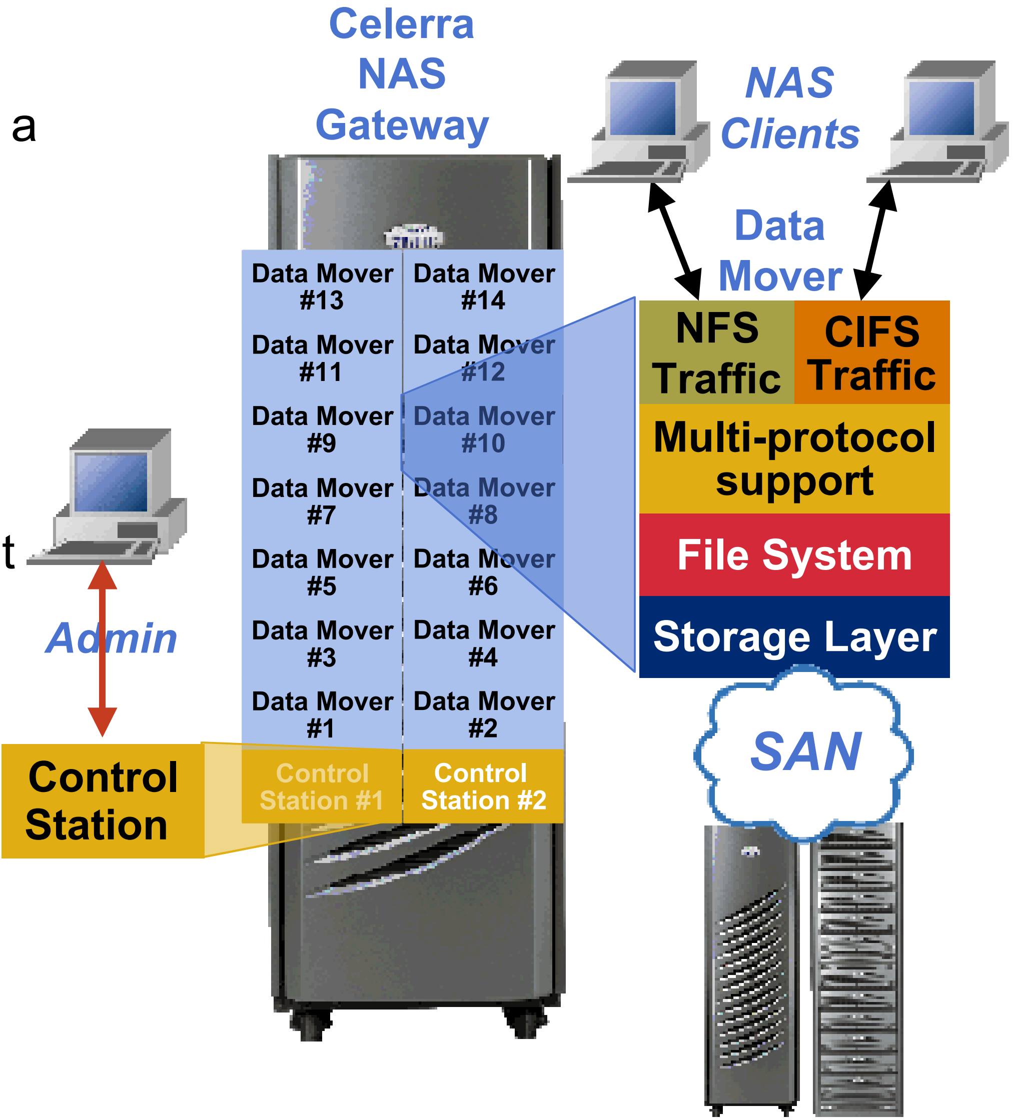
Highlights

- Scalable: 2 TB to 208 TB usable
- Integrated high availability
- Quick reboot offerings
- Fibre Channel and ATA Drives
- NAS and SAN integration



EMC Celerra

- Up to 14 file servers contained in a single clustered system
- Managed as a single server
- NAS front-end scales independently of SAN back-end
- N to 1 failover options
- Control Station
  - Administration & management
  - Web-based GUI
  - Manages Data Mover failover
- Data Mover
  - Optimized real-time OS
  - Concurrent NFS and CIFS file access
  - No performance impact after failover

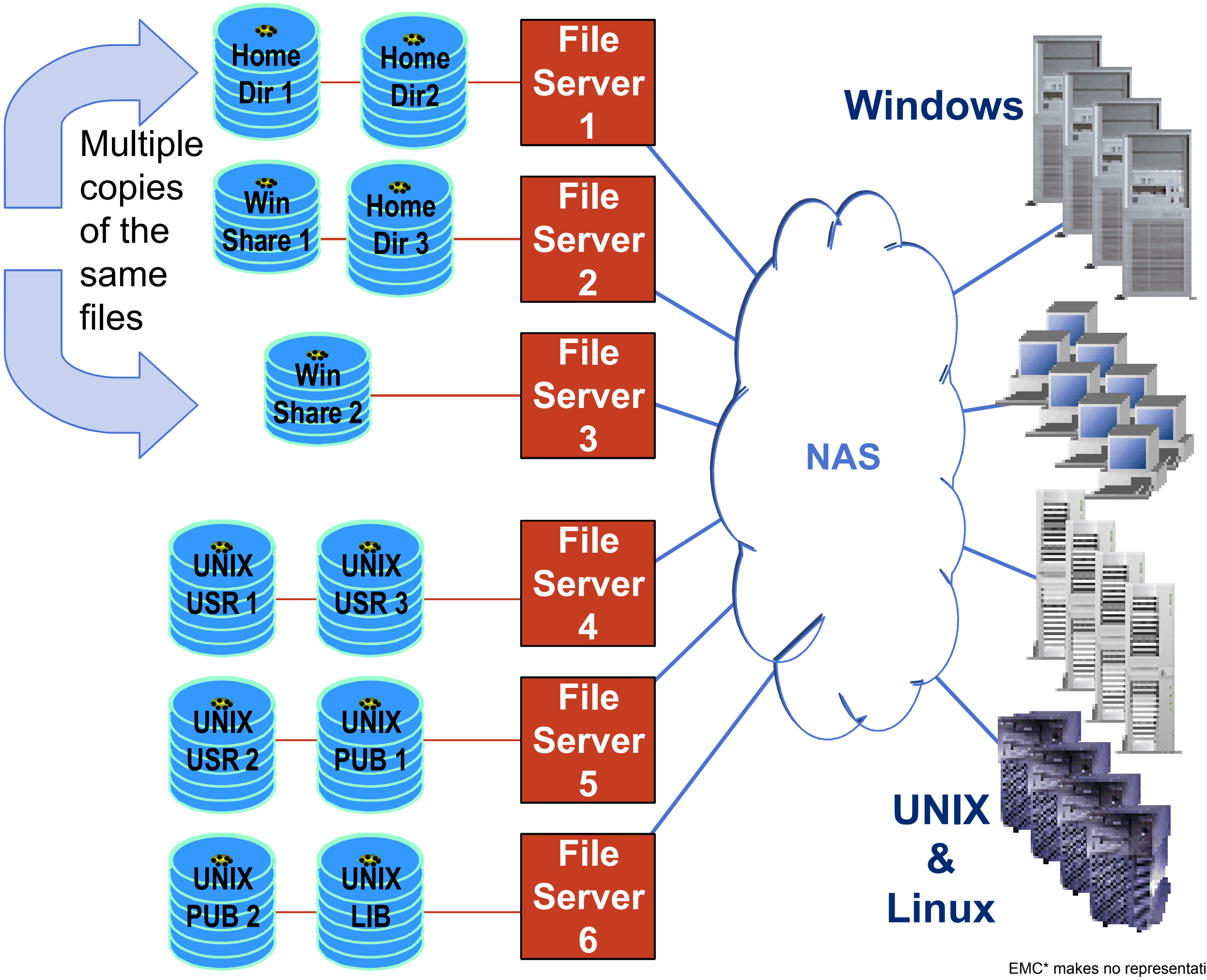


NAS Usage Scenarios

- Massive Consolidation Workloads
  - Cluster FS, Single Namespace
- Tiered Storage
  - Celerra FileMover API
- High Performance Computing
  - Multi-path IP SAN Filesystem
- Integrated Block and File
  - iSCSI Target and Initiator

EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

Massive Consolidation Workloads: Before



Client View

"FileServer1\HomeDir1"
"FileServer1\HomeDir2"
"FileServer2\HomeDir3"

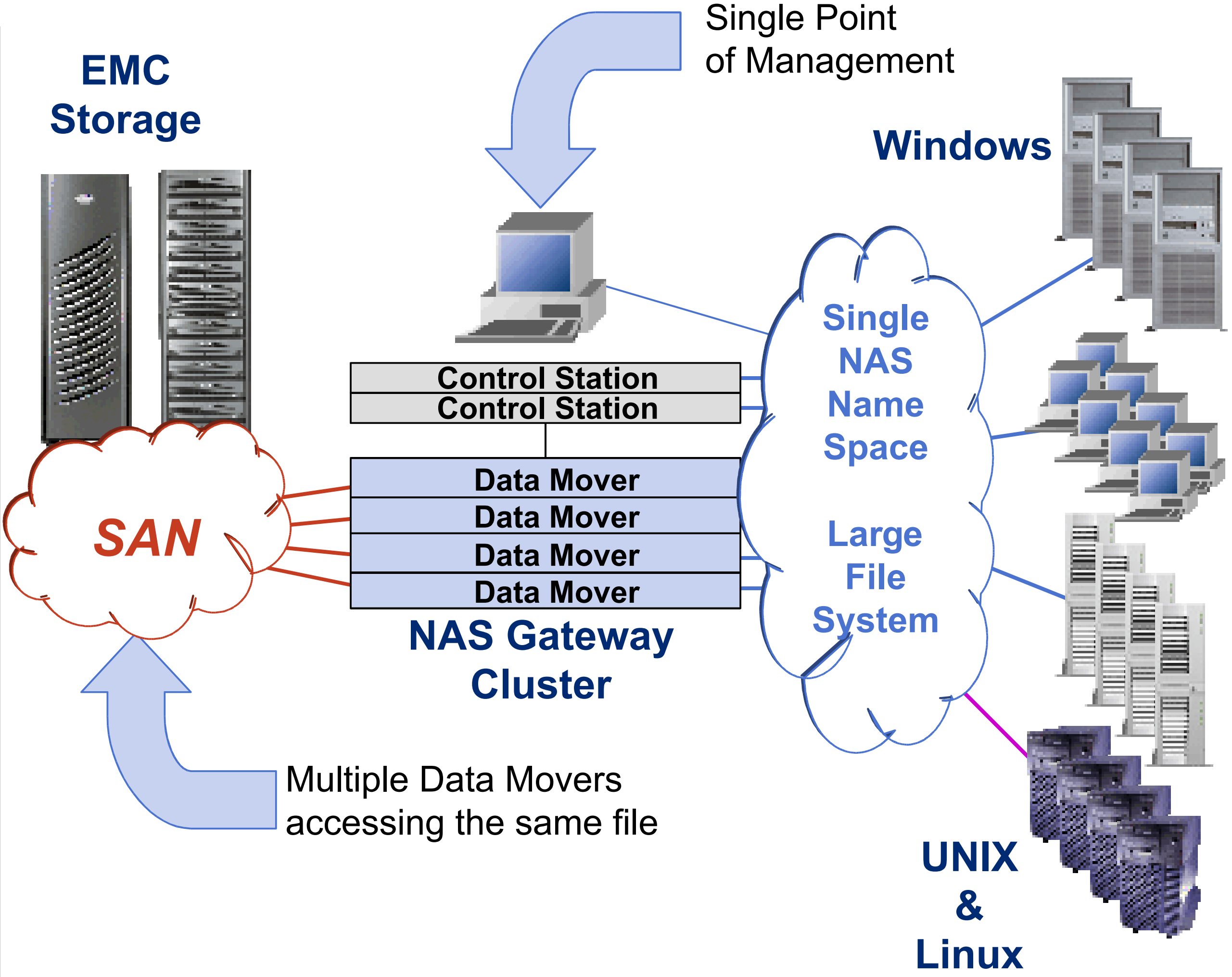
"FileServer2\WinShare1"
"FileServer3\WinShare2"

"FileServer4\USR1"
"FileServer3\USR2"
"FileServer4\USR3"

"FileServer5\PUB1"
"FileServer6\PUB2"
"FileServer6\LIB"

EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

Massive Consolidation Workloads: After



Single Point
of Management

EMC
Storage

Control Station

Control Station

Data Mover

Data Mover

Data Mover

Data Mover

NAS Gateway
Cluster

SAN

Multiple Data Movers
accessing the same file

Single
NAS
Name
Space

Large
File
System

Windows

UNIX
&
Linux

Client View

“Windows Home”

“Windows Share”

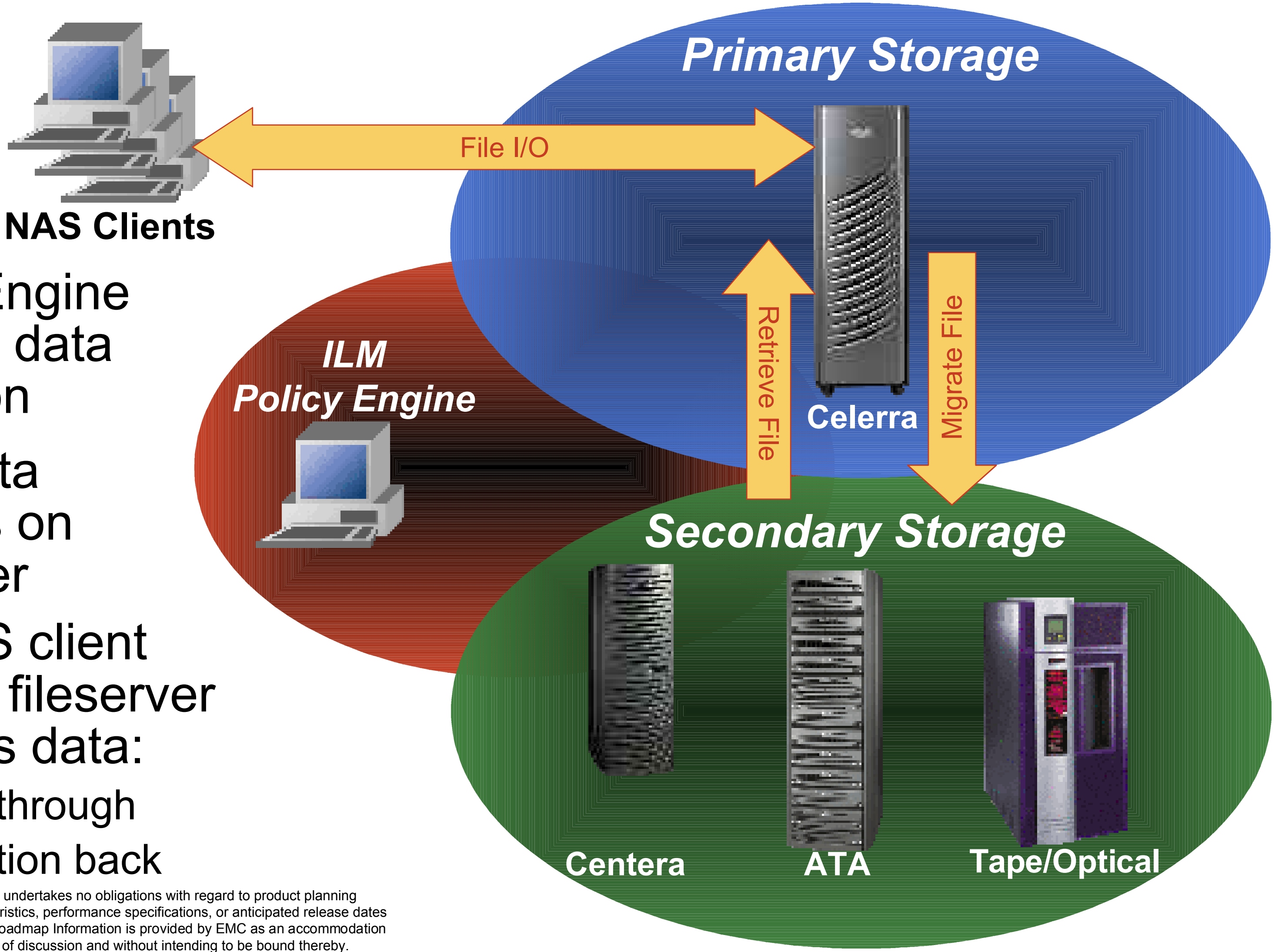
“UNIX USR”

“UNIX PUB”

”UNIX LIB”

EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, “Roadmap Information”). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

Tiered Storage: Celerra FileMover API

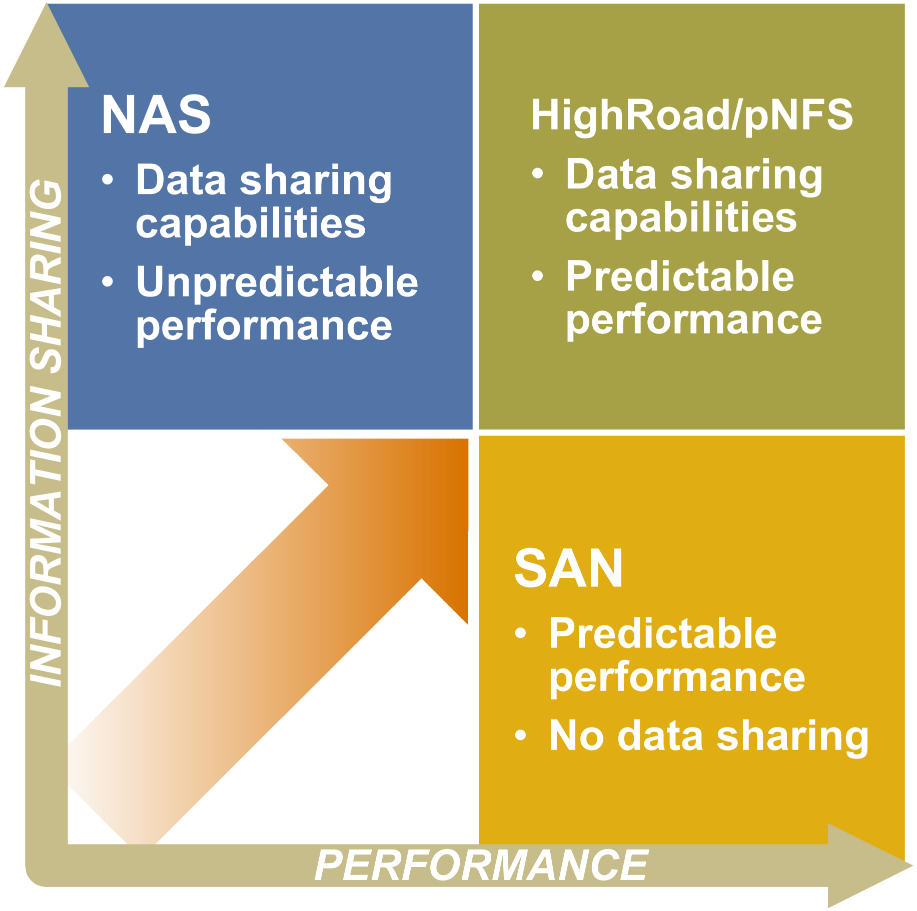


1. Policy Engine controls data migration
2. Metadata remains on fileserver
3. On NAS client access, fileserver retrieves data:
   - Pass through
   - Migration back

EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

October 12-14, 2004

Celerra HighRoad and Parallel NFS (pNFS)



NAS
- Data sharing capabilities
- Unpredictable performance

HighRoad/pNFS
- Data sharing capabilities
- Predictable performance

SAN
- Predictable performance
- No data sharing

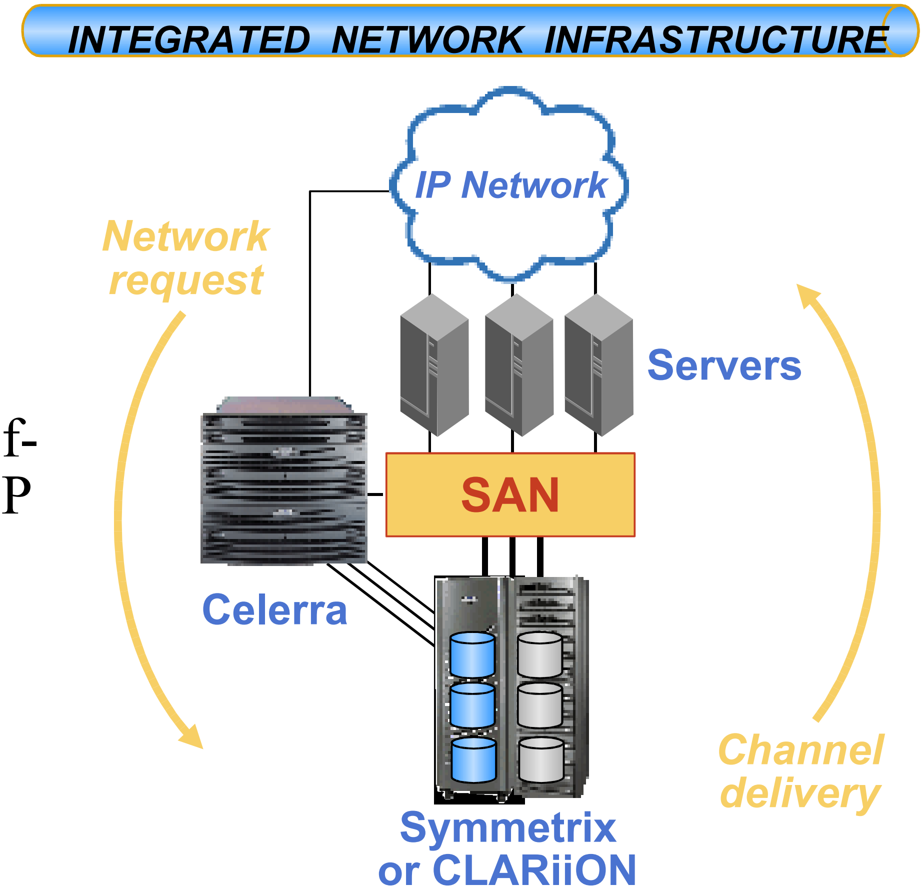
Applications

- Media
  - Post production
  - Television finishing
  - Streaming video
  - Advertising
- Large image processing
  - Seismic
  - Medical
  - CAD / CAM
  - Scientific simulations
- Backup

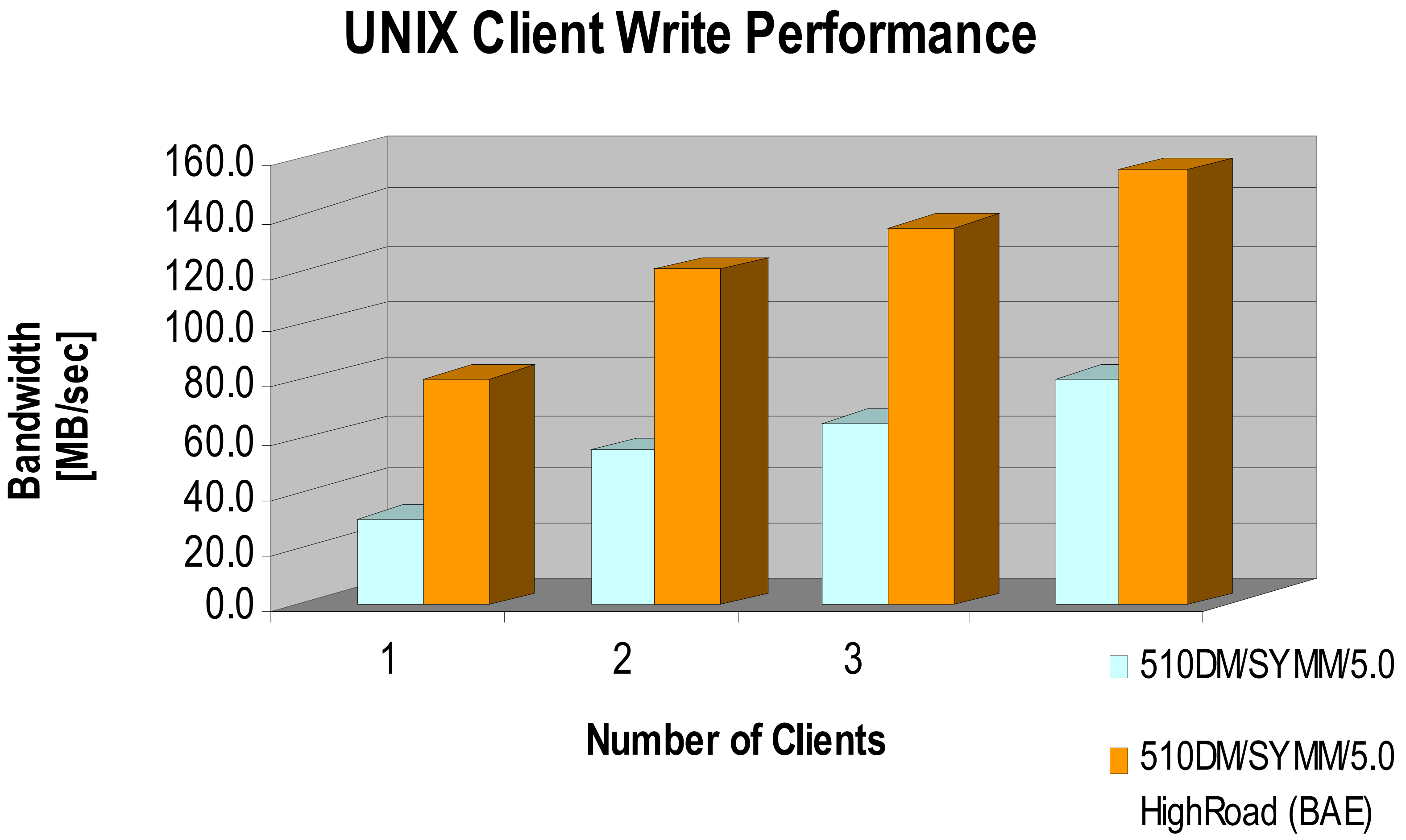
Address both sharing and bandwidth challenges

Celerra HighRoad and pNFS

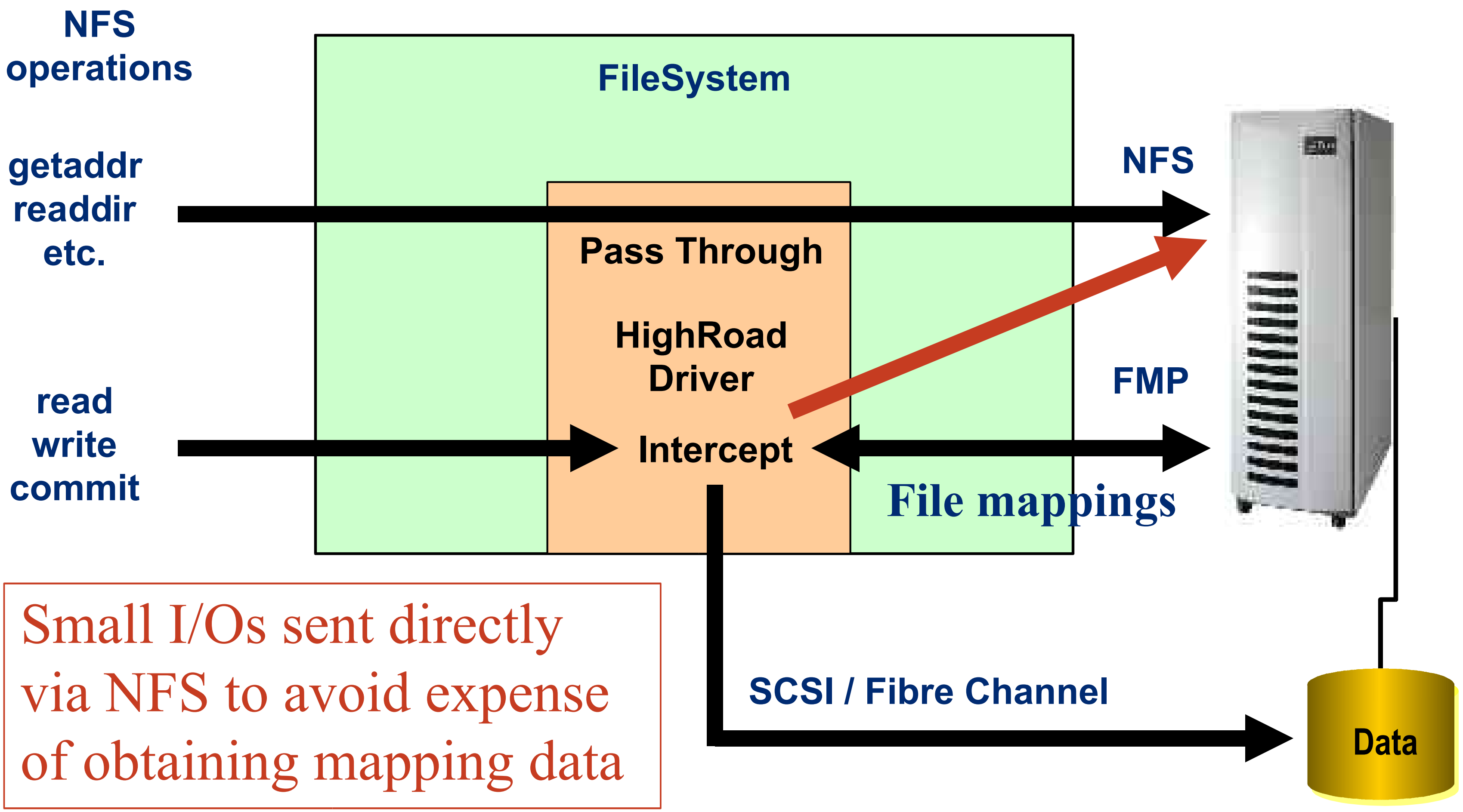
- Network Request and Channel Delivery
  - Servers connected to storage over SAN
  - Servers connected to an out-of-band “meta data” cluster via IP
  - Servers send file requests to cluster via NFS/CIFS
  - Data access is direct via SAN (performance)
  - Meta data cluster scaling improved by data bypass



HighRoad Performance



HighRoad Client Driver



NFS operations

getaddr
readdir
etc.

read
write
commit

FileSystem

Pass Through

HighRoad Driver

Intercept

NFS

FMP

File mappings

SCSI / Fibre Channel

Data

Small I/Os sent directly via NFS to avoid expense of obtaining mapping data

HighRoad Metadata protocol: FMP

- Client asks fileserver “where is this file?”
- Fileserver provides map as answer
  - And grants read/write access permissions to client
  - Client can now read/write file blocks (via SAN)
- Update server state as needed
  - Hole filling (zeros or data)
    - Client does writes, tells server what it did
  - File extension or truncation (new EOF)
    - Server propagates new EOF to other clients

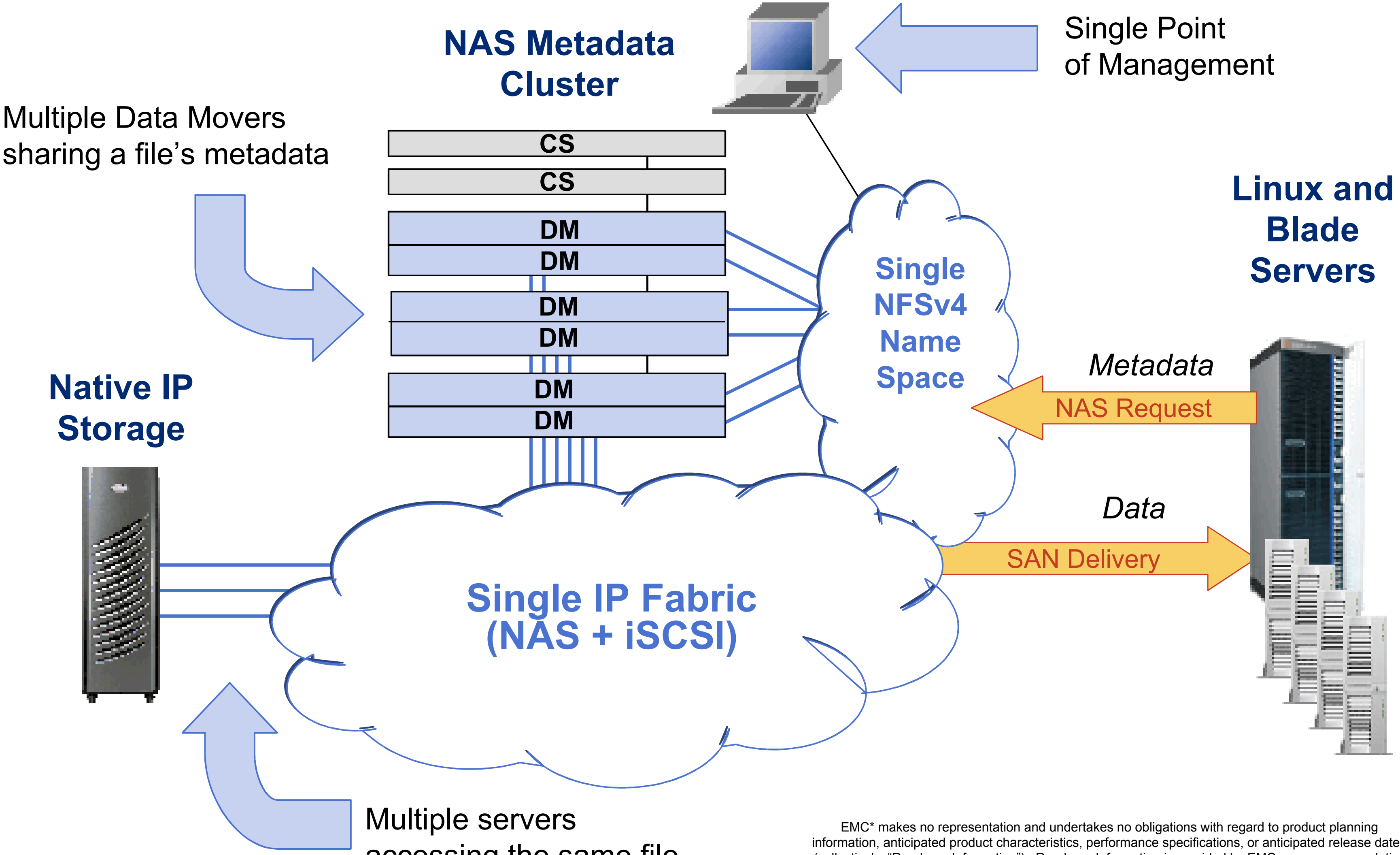
FMP and pNFS

- FMP: Stand alone protocol
  - Can be used with NFSv2, NFSv3, NFSv4 and CIFS
  - Data caching and consistency are independent of protocol
- pNFS: NFSv4 extensions for mapping
  - Allow use of compound operations
  - Similar functional behavior to FMP
  - To be standardized in IETF
- FMP → pNFS
  - EMC has provided FMP specification to the pNFS effort
  - Enable pNFS effort to learn from HighRoad experience
  - Celerra HighRoad product will evolve to support pNFS

Some Lessons Learned from HighRoad

- Volume identification (where’s the file/FS?)
  - Have to use volume and/or filesystem labels
  - Addresses don’t work (what’s a SCSI address?)
- Access permission recall is essential and subtle
  - Server recall may conflict with pending client request
  - Out-of-order delivery can create race conditions
- Block permission granularity makes a difference
  - Whole file granularity creates false sharing conflicts
- Keep-alive needed to detect client death
  - And clean up any access permissions it held

High Performance Computing Workloads



NAS Metadata Cluster

Single Point of Management

Multiple Data Movers sharing a file's metadata

CS

CS

DM

DM

DM

DM

DM

DM

Single NFSv4 Name Space

Linux and Blade Servers

Metadata

NAS Request

Data

SAN Delivery

Native IP Storage

Single IP Fabric (NAS + iSCSI)

Multiple servers accessing the same file

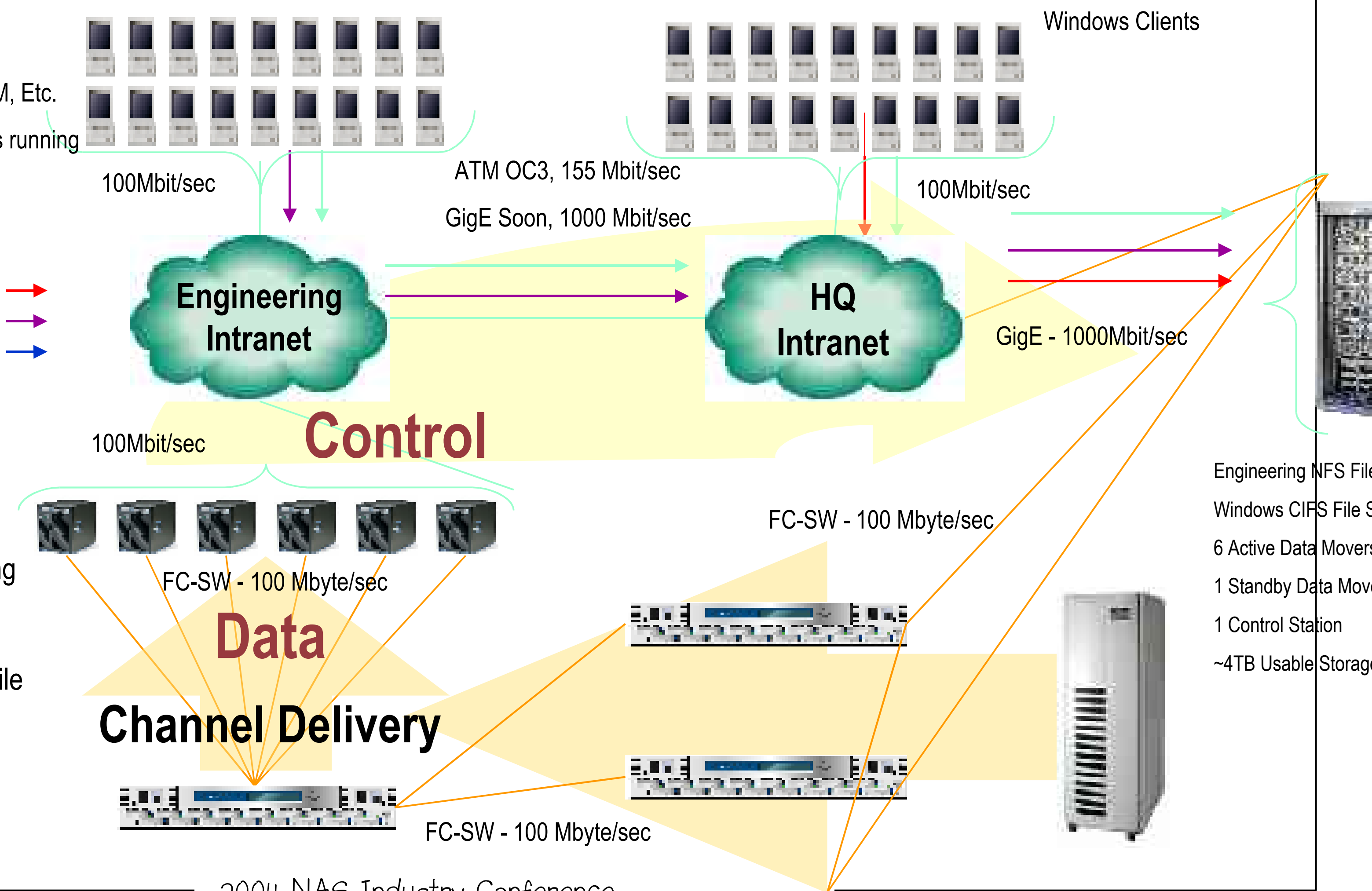
EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

October 12-14, 2004

Example: Consumer Electronics Manufacturer

- 150-200 UNIX workstations
- Sun, SGI, HP, IBM, Etc.
- Engineering users running CAD/CAM/CAE

6 HP J6000 for Engineering Simulation runs on LSF Computer Engines. File sharing at channel speeds. 100Mbyte/sec data access to common file system.



Engineering NFS File
Windows CIFS File S
6 Active Data Movers
1 Standby Data Move
1 Control Station
~4TB Usable Storag

🟥 CIFS Data Flow
🟪 NFS Data Flow
🟦 FTP Data Flow

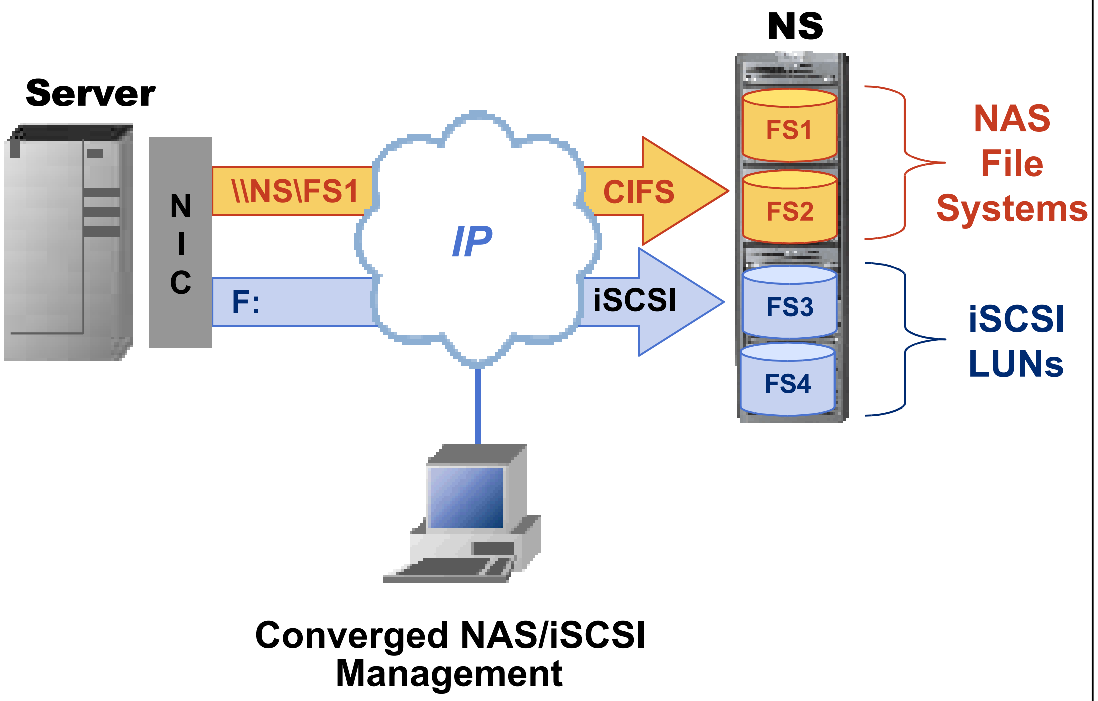
2004 NAS Industry Conference

October 12-14, 2004

Block and File Workloads

iSCSI Target

- Microsoft Logo Certified
  - iSNS Naming Service
  - CHAP Authentication
- Simple Management
  - Web-based GUI
  - Common toolset for NAS and IP SAN
- High Availability
  - Data Mover failover
  - Port/path failover

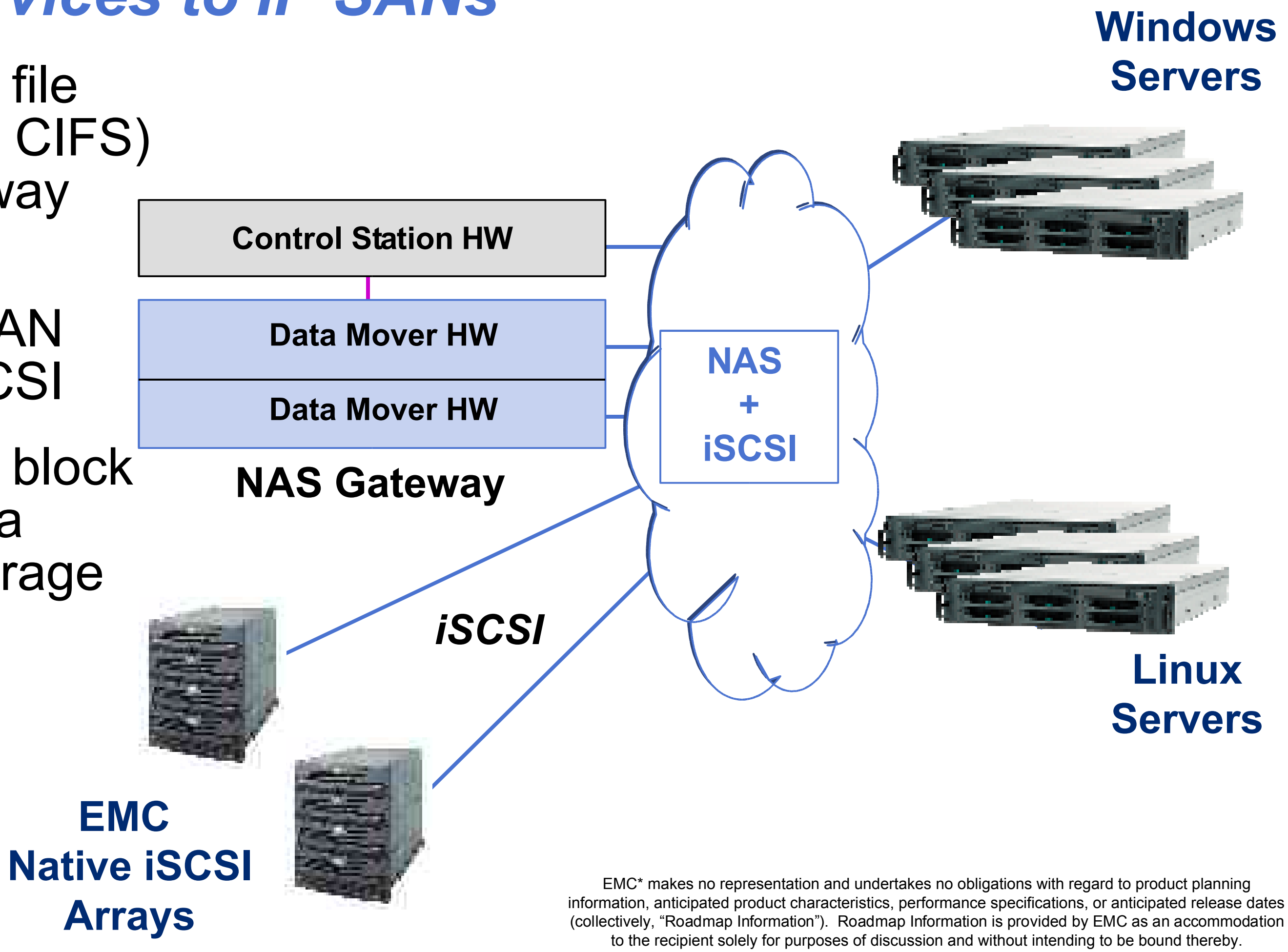


EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

NAS Gateway iSCSI Initiator

Adds NAS Services to IP SANs

- Clients access file data (NFS and CIFS) via NAS Gateway
- NAS Gateway accesses IP SAN storage via iSCSI
- Clients access block data directly via iSCSI to IP storage

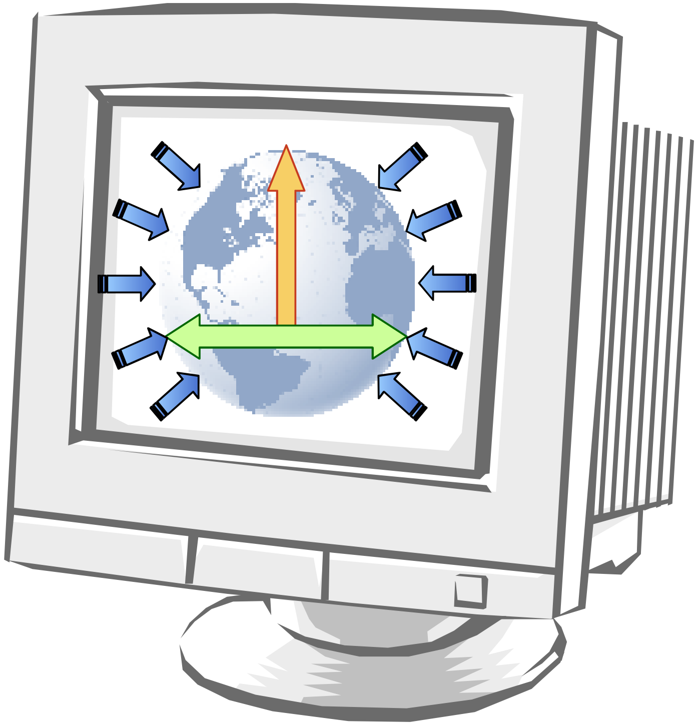


EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

Delivering on the EMC NAS Vision

- Infinite Scalability
  - Cluster File System and Large File Systems
  - Multi-path IP SAN File System
- Optimized Data Placement
  - FileMover API
- Global Accessibility
  - Single Name Space
- Centralized Management
  - Single Name Space
  - ISCSI Target and Initiator

Delivering on ILM



EMC* makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, “Roadmap Information”). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

EMC^2

where information lives