Emerging Technology: Scalable Storage Clustering

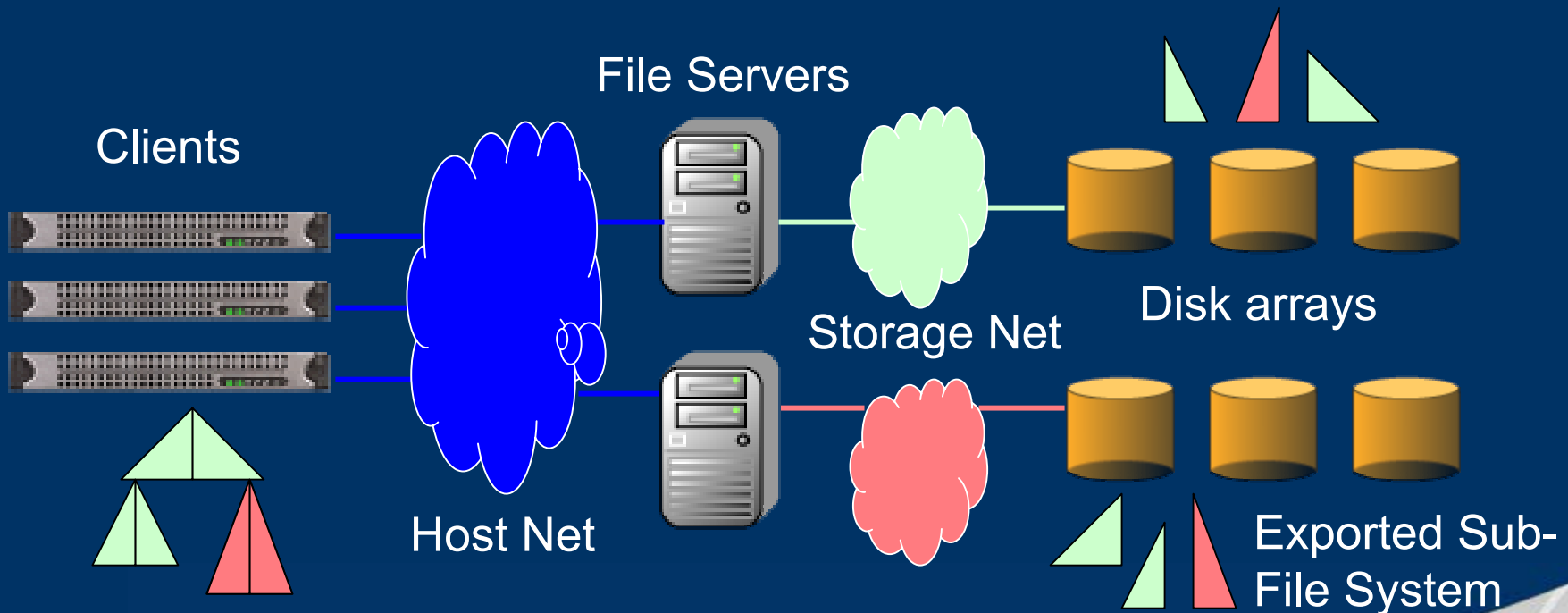
Garth Gibson
garth@panasas.com

Panasas Inc, and
Carnegie Mellon University

NASA/IEEE MSST 2004 April 15, 2004

Today's NAS Clustering Practices

- **Customer namespace decomposition**
  - Expensive administrator given messy tasks
  - Single files & directory decomposition very visible to users



Clusters Demand a New Storage Architecture

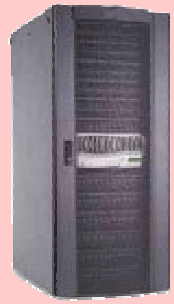
Traditional HPC Computing

Monolithic Computers

BOTTLENECK

A few data paths

Monolithic Storage



Cluster Computing

Linux Compute Cluster

?

More cost-effective boxes

Scale:
file & total bandwidth
file & total capacity
load & capacity balancing

With lower $ / Gbps

Scalable Storage Cluster Architecture

Lesson of compute clusters: Scale out commodity components

- **Bladeserver approach provides**
  - High volumetric density
  - Incremental growth, pay-as-you-grow model
  - Needs single system image SW architecture

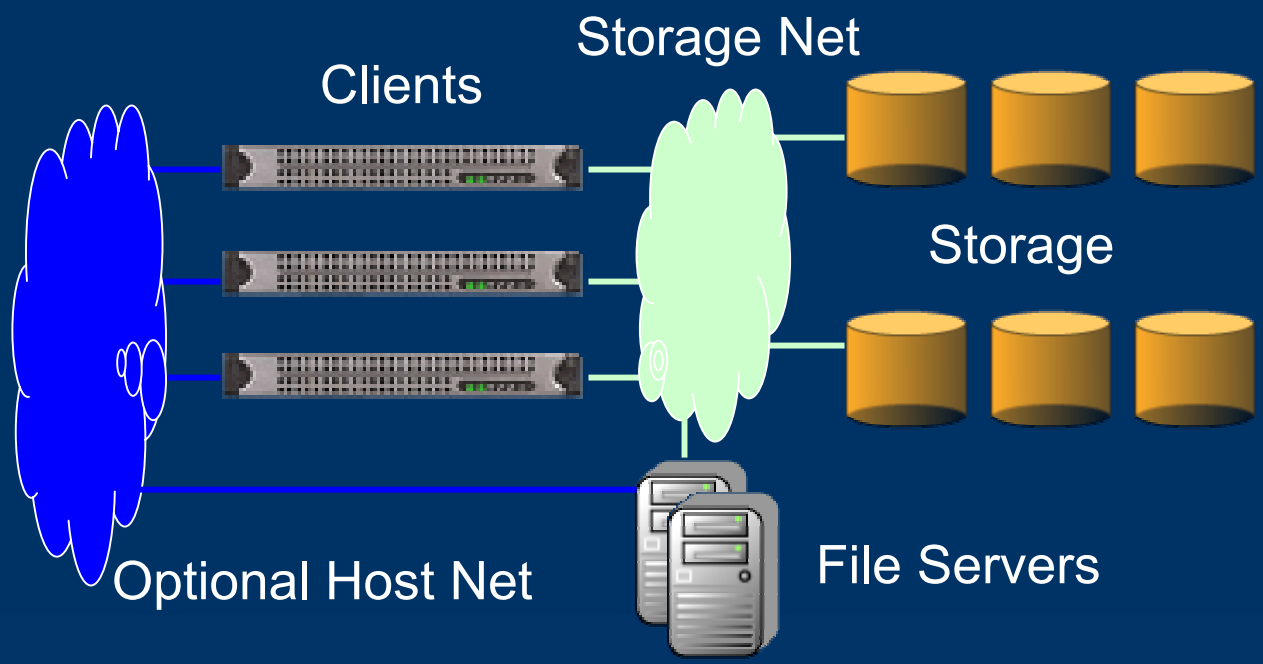
StorageBlade
2 SATA spindles

Shelf of Blades
5 TB, 4 Gbps

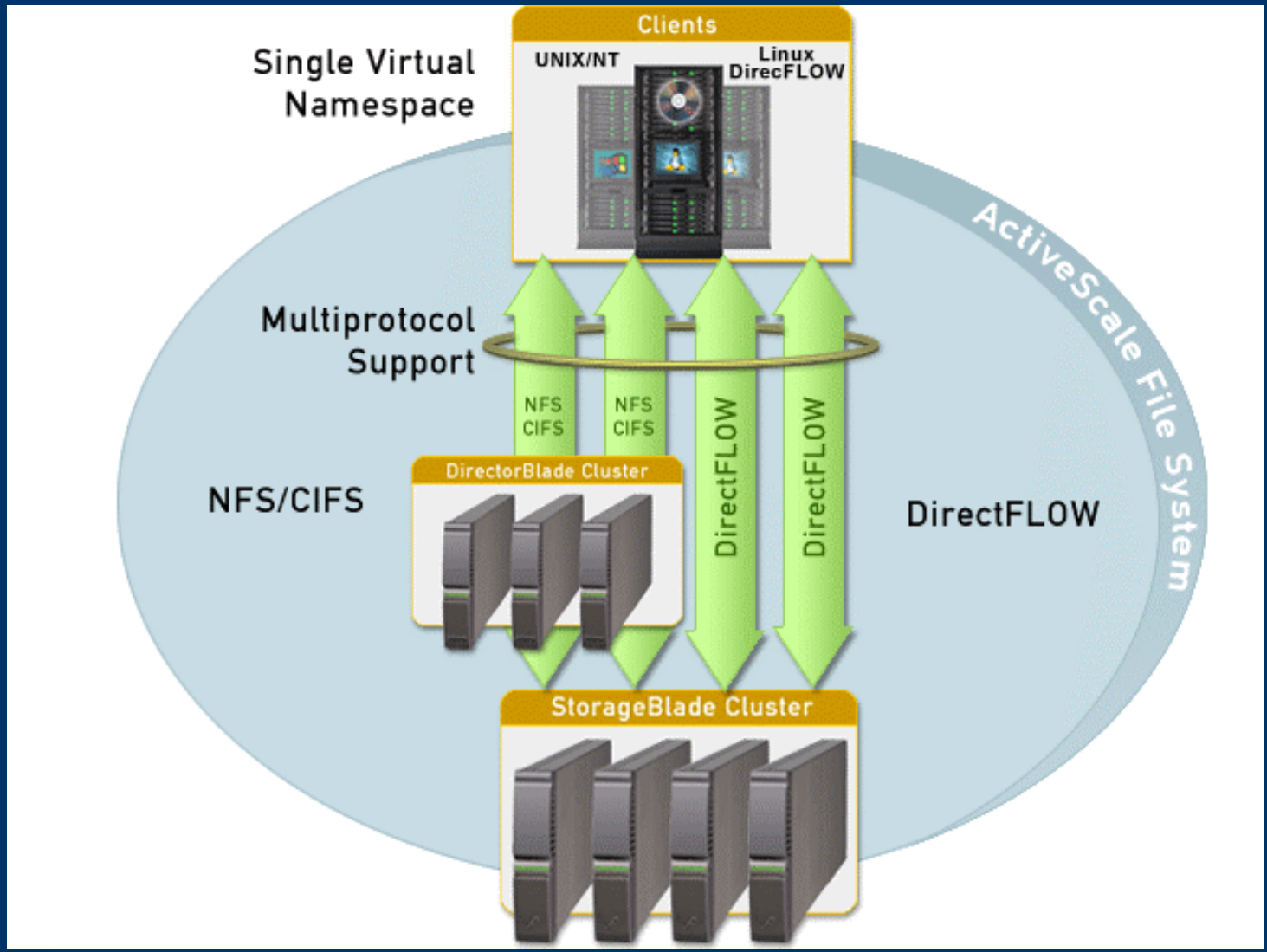
Single System Image
55 TB, 44 Gbps per rack

Scale Out File Service w/ Out-of-Band

- **Client sees many storage addresses, accesses in parallel**
  - Zero file servers in data path
  - Examples: IBM SAN FS, EMC High Road, SGI CXFS, Panasas, etc
  - Mostly built on block-based SANs where ***servers trust all clients***

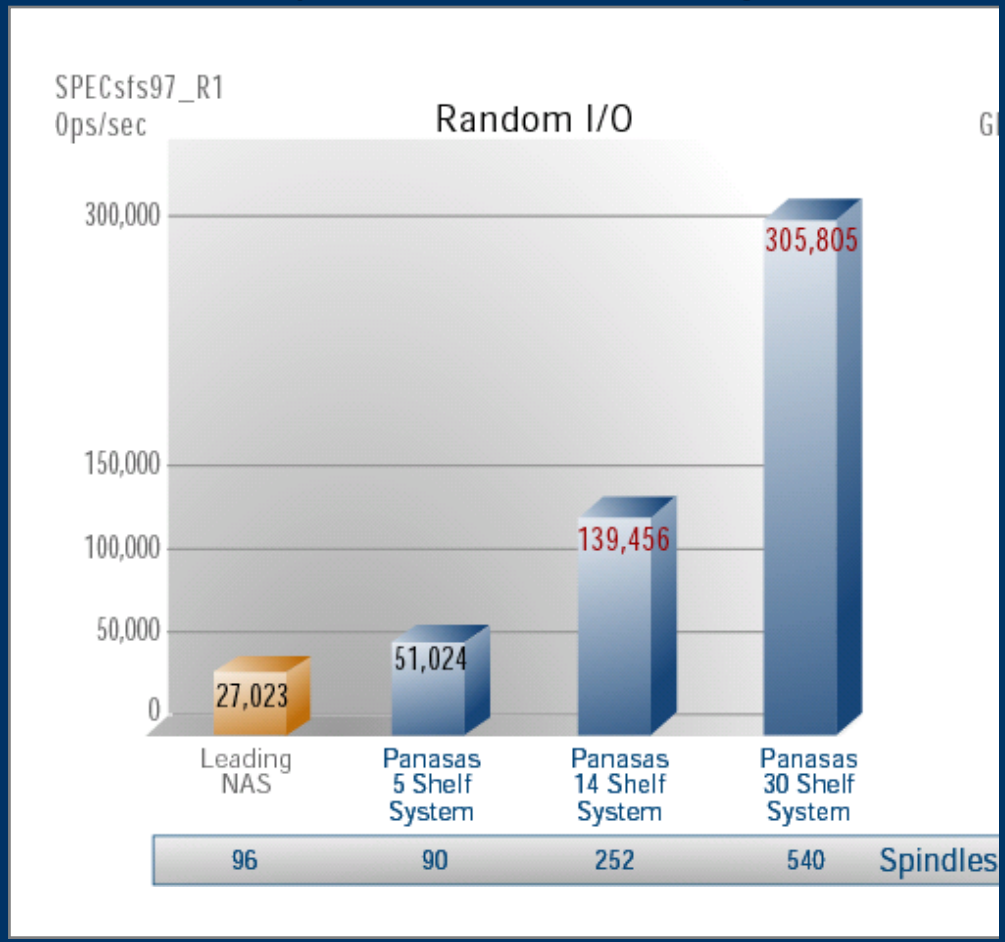


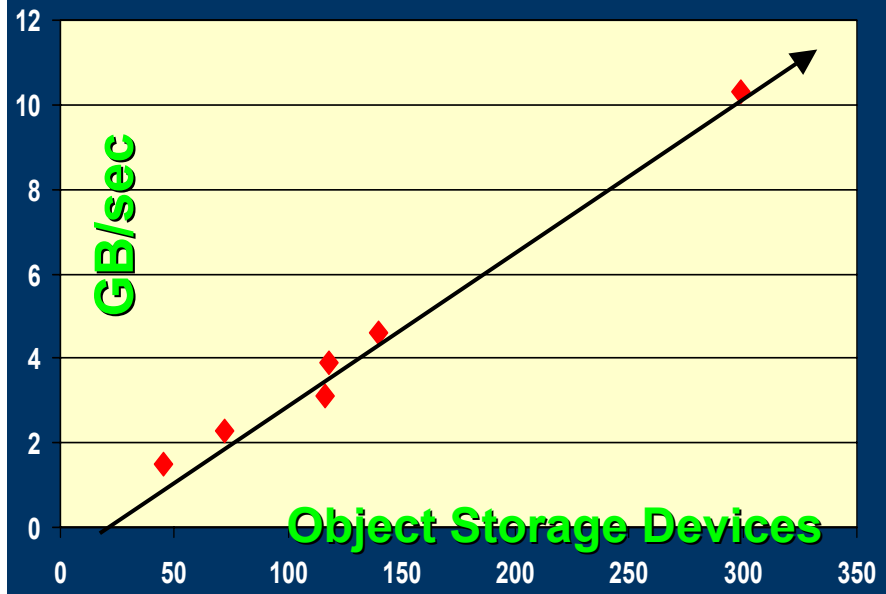
Asymmetric Out-of-band & Clustered NAS



Performance & Scalability for all Workloads

Objects: breakthrough data throughput AND random I/O



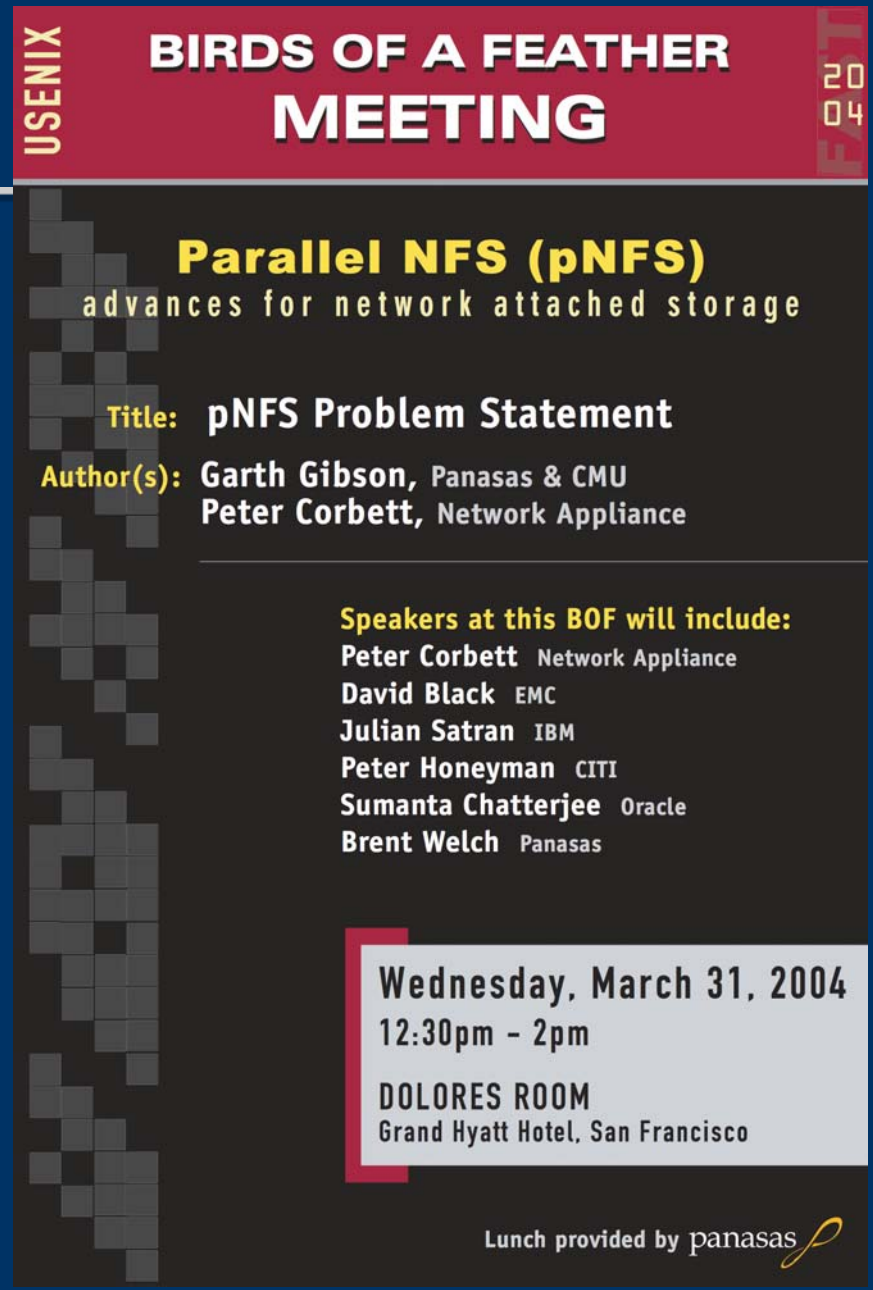


Source: SPEC.org & Panasas

pNFS

- **December 2003 U. Michigan workshop**
  - Whitepapers at www.citi.umich.edu/NEPS
  - U. Michigan CITI, EMC, IBM, Johns Hopkins, Network Appliance, Panasas, Spinnaker Networks, Veritas, ZForce
- **NFSv4 extensions enabling clients to access data in parallel storage devices of multiple standard flavors**
  - Parallel NFS request routing / file virtualization
  - Extend block or object "layout" information to clients, enabling parallel direct access
- **Problem Statement Internet-Draft published**

  draft-gibson-pnfs-problem-statement-00.txt



panasas

Out-of-band Value Proposition

- Out-of-band means client uses more than one storage address for a given file, directory or closely linked set of files

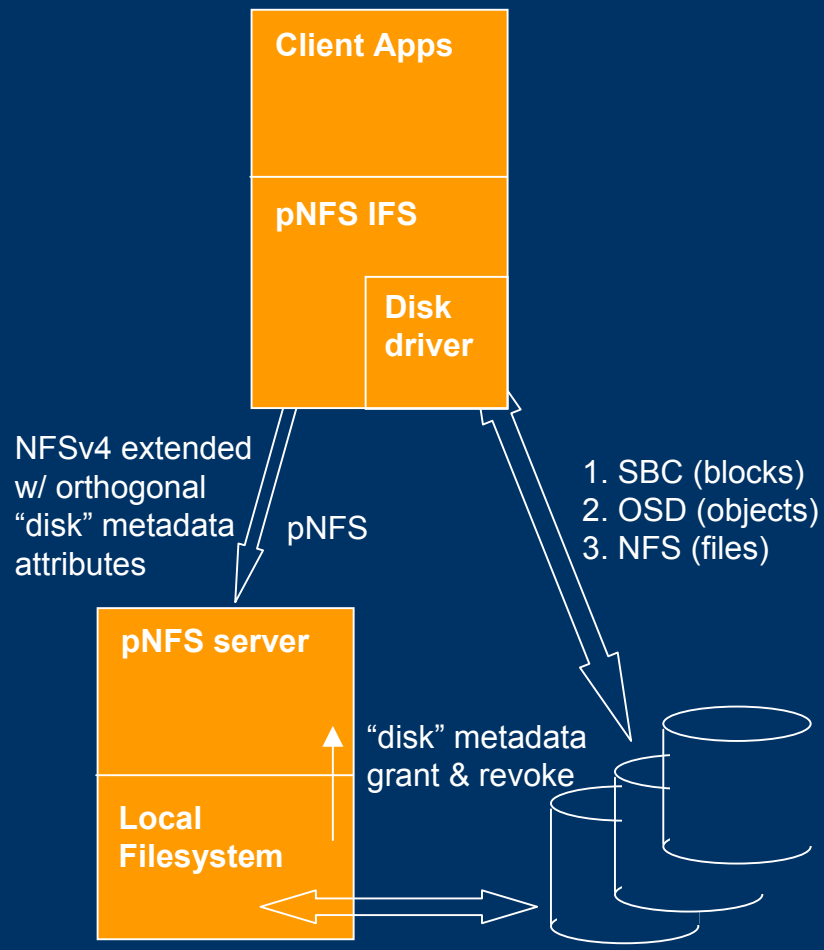
- **Scalable capacity:** file/dir uses space on all storage: can get big
- **Capacity balancing:** file/dir uses space on all storage: evenly
- **Load balancing:** dynamic access to file/dir over all storage: evenly
- **Scalable bandwidth:** dynamic access to file/dir over all storage: big
- **Lower latency under load:** no bottleneck developing deep queues
- **Cost-effectiveness at scale:** use streamlined storage servers
- If wire standards lead to **standard client SW:** share client support $$$

NFS Extensions Approach

- **Limited (Market)/(support $) for proprietary advanced FS client SW**
  - Customers: interoperating competition; vendors: too many changing client platforms
- **Rally behind one open industry-standard advanced FS client SW**
- **IETF NFS is unrivalled as open industry-standard FS client SW**
  - Raising (Market)/(support $) is worth giving up proprietary feature control
- **V4: "Recallable delegations allow clients holding a delegation to locally make many decisions normally made by the server"**
- **Propose a new flavor of NFSv4 delegations**
  - A client requesting a delegation asks for out-of-band file address maps
  - Server protects map integrity while delegation lasts, knowing file data may change
  - Server can re-synch with file contents by recalling the delegations

Multiple Data Server Protocols

- **Inclusiveness favors success**
- **Three (or more) flavors of out-of-band metadata attributes:**
  - BLOCKS:
    SBC/FCP/FC or SBC/iSCSI…
    for files built on blocks
  - OBJECTS:
    OSD/iSCSI/TCP/IP/GE for files built on objects
  - FILES:
    NFS/ONCRPC/TCP/IP/GE for files built on subfiles
- **Inode-level encapsulation in server and client code**



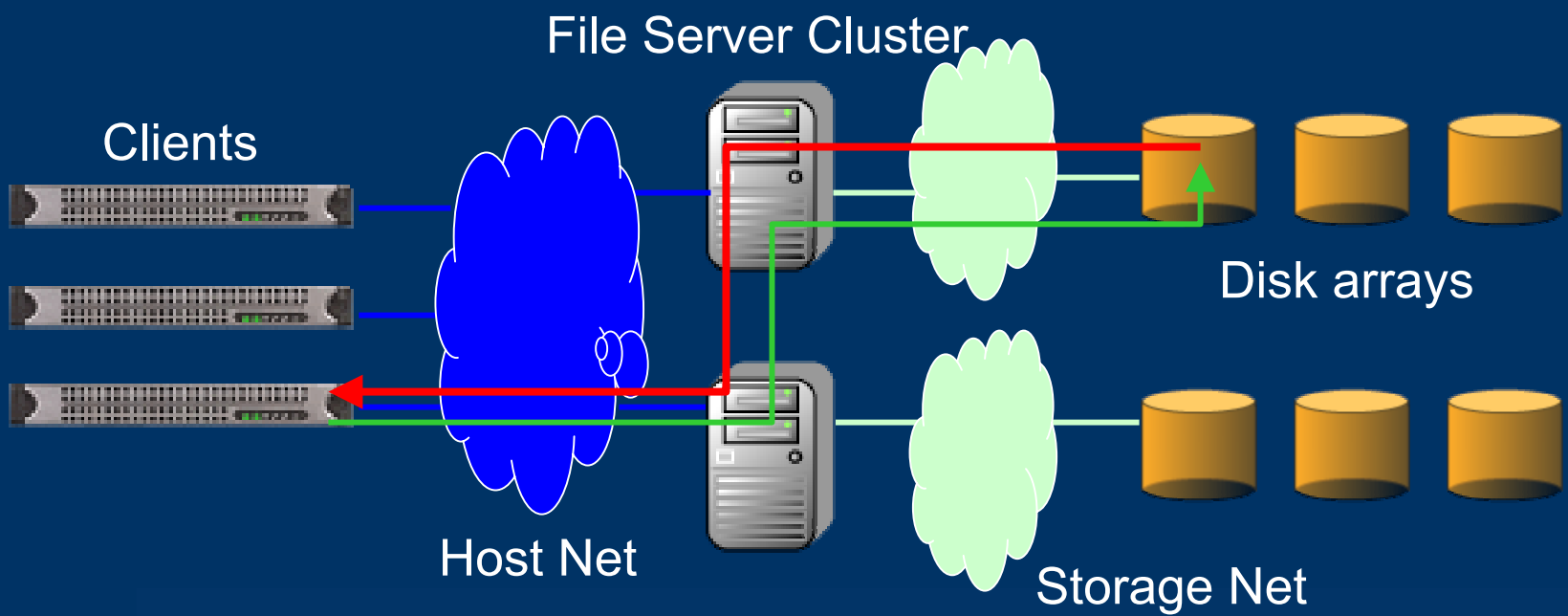
Cluster storage for scalable Linux clusters

Garth Gibson
ggibson@panasas.com
www.panasas.com

NASA/IEEE MSST 2004 April 15, 2004

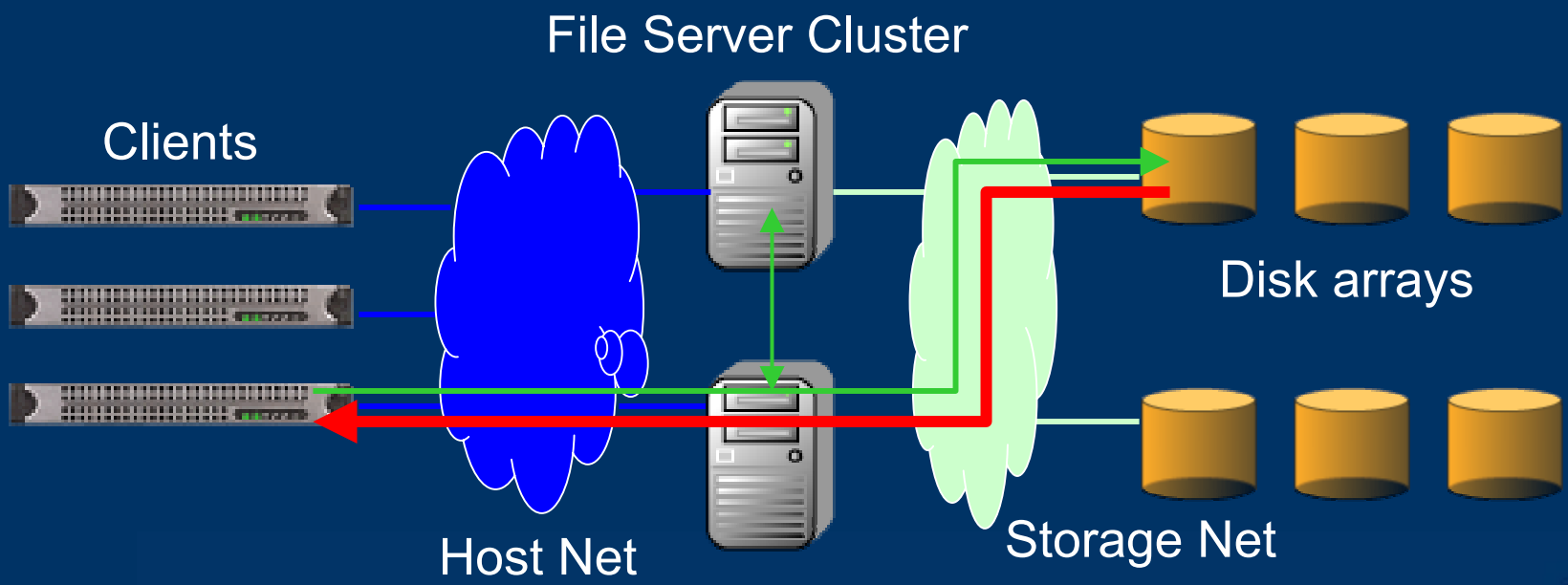
Scale Out Phase 2: Forwarding Servers

- **Single system image with forwarding**
  - Mount point binding less relevant
  - Control and data pass through two servers
  - Single file system bandwidth limited by backend server & storage



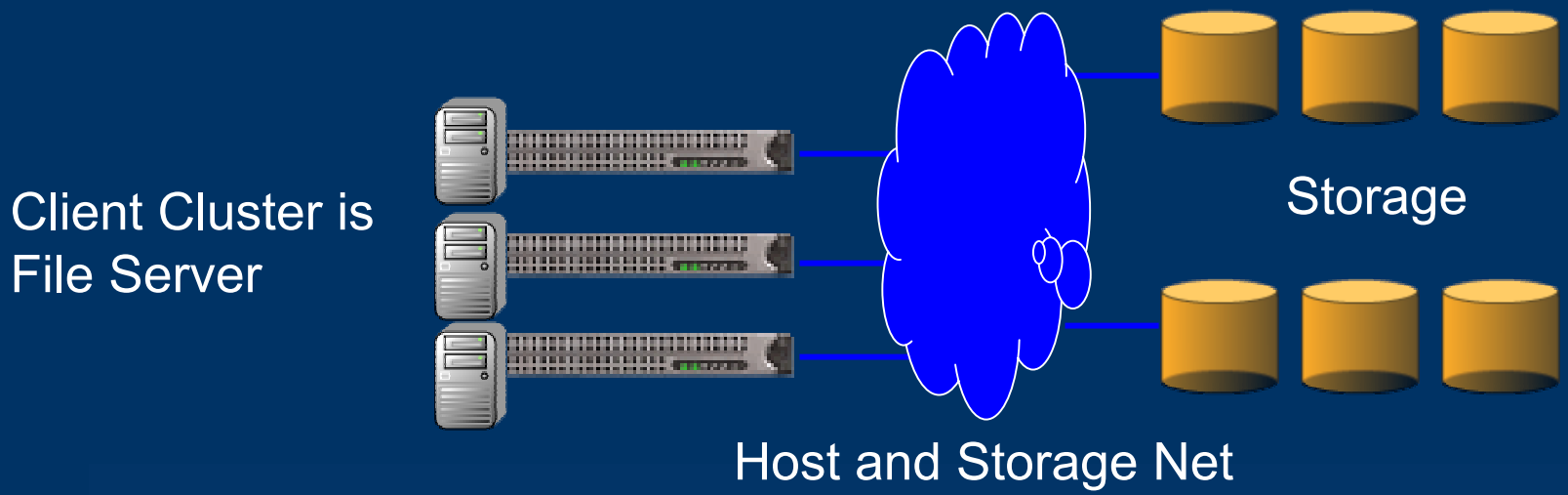
Scale Out Phase 3: Any Server Will Do

- **Single server does all data transfer in single system image**
  - Servers “hand off” role of accessing storage
  - Control and data pass through one server
  - Single file system capacity & bandwidth scales with server cluster



Scale Out Phase 4S: Symmetric OOB

- **Every client is fully capable server**
  - Client's local code acquires locks & metadata from other clients
  - Examples: RedHat GFS, IBM GPFS, Sun QFS etc
  - Trust boundary includes all clients



New *Object Storage* Architecture

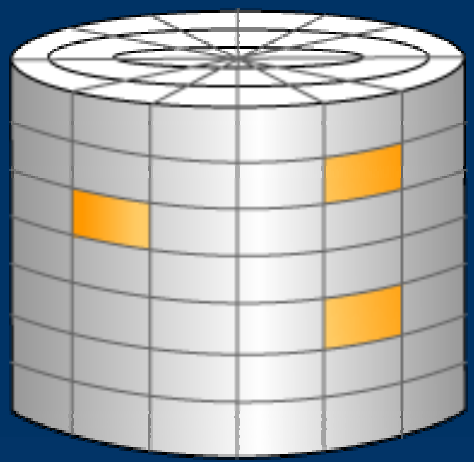
- **An evolutionary improvement to standard SCSI storage interface**
- **Raises level of abstraction: Object is container for "related" data**
  - Storage *understands* how different blocks of a "file" are related
- **Offload most datapath work from server to intelligent storage**

Block Based Disk

Operations:
Read block
Write block

Addressing:
Block range

Allocation:
External

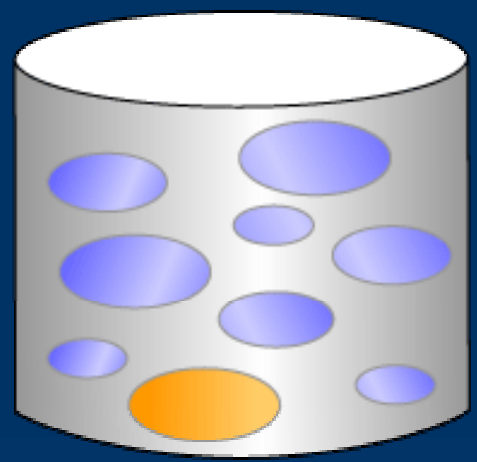


Object Based Disk

Operations:
Create object
Delete object
Read object
Write object

Addressing:
[object, byte range]

Allocation:
Internal



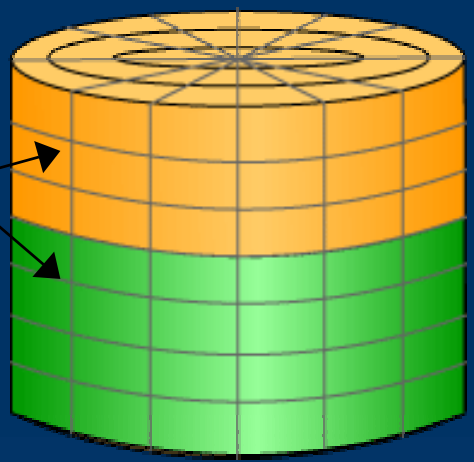
Source: Intel

Additional Strengths Of Object Model

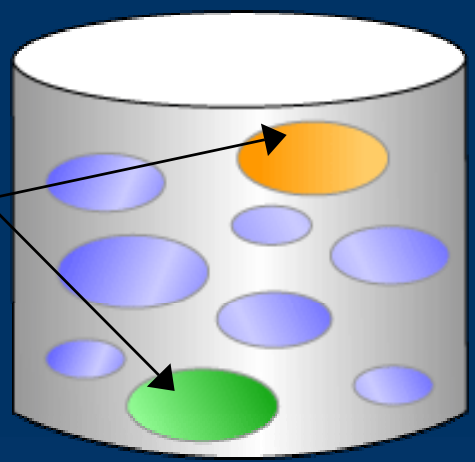
Disk knows relationships among data within object

- **Allows autonomous action within Object Storage Device**
  - Storage evolves to become self-managing, optimized to workload of each file
- **Rich sharing among heterogeneous clients via higher level interface**
- **Finer granularity of security: protect & manage one file at a time**

Security At Volume Level

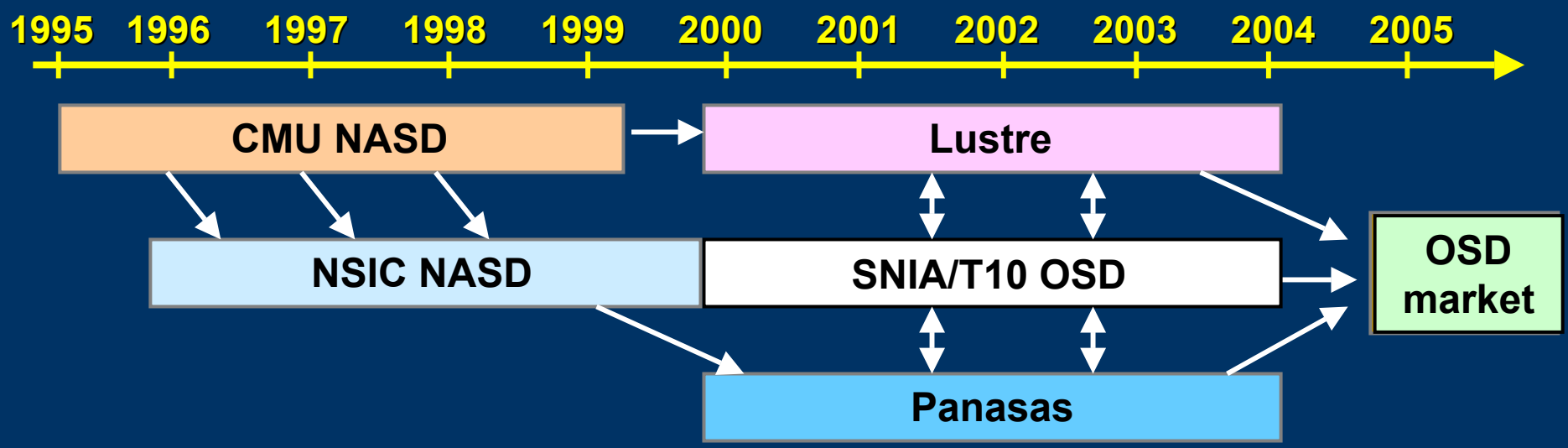


Security At Object Level



Source: Intel

Standardization Timeline



- **SNIA TWG has issued OSD protocol to T10 standards body**
  - Co-chaired by IBM and Intel, protocol is a general framework (transport independent)
  - Sub-committee leadership includes IBM, Seagate, Panasas, HP, Veritas, ENDL
  - Product commitment from HP/Lustre & Panasas; research projects at IBM, Seagate
  - T10 letter ballot (closed Mar 24 04) passed; integrating comments for full public review
  - www.snia.org/tech_activities/workgroups/osd & www.t10.org/ftp/t10/drafts/osd/osd-r09.pdf

Object Storage Systems

Expect wide variety of Object Storage Devices

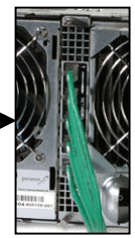
- Disk array subsystem
- Ie. LLNL with Lustre

- "Smart" disk for objects
- 2 SATA disks – 240/500 GB

- Prototype Seagate OSD
- Highly integrated, single disk

- Orchestrates system activity
- Balances objects across OSDs

- **Stores up to 5 TBs per shelf**



16-Port GE Switch Blade

- 4 Gbps per shelf to cluster

How Does an Object Scale?

Scale capacity, bandwidth, reliability by striping according to small map

File
Comprised of:
- User Data
- Attributes
- Layout