New Approaches to Supercomputing with Panasas

Takahiko Tomuro
Technical Consultant for Panasas
Scalable Systems Co., Ltd.
tomuro@sstc.co.jp

Scalable Systems Co., Ltd.

1986 Cray Research Japan,Ltd.
The company is led on an activity and technological sides of SE, the sales support, and the marketing support, etc.

1996 SGI Japan Ltd.
SE director, VP of Product & Marketing

2003 Chief Technology Officer
The introduction to the customer and the activity of alliance with each company of a wide-ranging technological trend to say nothing of the SGI product were done.

2005 Scalable Systems Co.. Ltd Ltd.

SSTC High Performance and Productivity Scalable Systems Co.,Ltd.

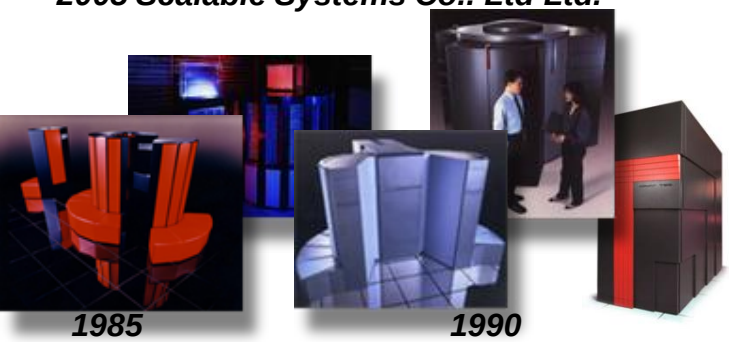
2005

Scalable Systems

Scalable Systems makes the best use of the experience related to abundant HPC in CRAY and SGI, and offers a new solution.

1996 *2000* *2005*

Silicon Graphics



1985 *1990* *1995*

CRAY Research Inc.

It has acted for the offer of the HPC solution by systems of various architectures such as a vector computer, MPP systems, and super-servers (SUN compatible machine). It introduces a state-of-the-art technology to Japan by the vector processing and the parallel processing.

The HPC solution has been offered by first DSM (distributed shared memory system) and a large-scale NUMA system.

The commercialization of a scalable system by Linux and the Intel processor and the introduction support of the system.

Panasas Company Overview

- Company: Silicon Valley-based; venture-backed; 150 people WW
  - FEB 28: announced 105% rev growth for 2006 and investments in Asia
- Technology: Parallel cluster storage solutions for Linux HPC
- History: Founded 1999 by Garth Gibson, co-inventor of RAID
- Extensive HPC industry validation:

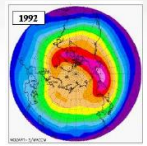
- Panasas parallel I/O and Storage Enabling Petascale Computing
  - Storage system selected for LANL's $110MM "Roadrunner" project
    - Petaflop IBM system with 16,000 AMD cpus + 16,000 IBM cells, 4x over LLNL BG/L
  - SciDAC selected Panasas CTO Garth Gibson to lead petascae PDSI
  - Panasas and Garth Gibson primary contributors to pNFS extensions

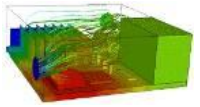
Panasas HPC Application Focus

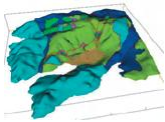
HPC Application and Sample Vertical

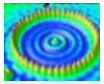
- Climate, Weather, and Ocean (CWO) Modeling
  - National Labs, Defense, Gov Agencies, Academia
- Computer-Aided Engineering (MCAE)
  - Automotive, Aerospace, Defense, Manufacturing
- Electronic Design Automation (EDA) and ECAE
  - Semi-conductor, IC Design, Systems
- Seismic Processing, Interpretation, Reservoir Sim
  - Energy Exploration and Production, Oil & Gas
- Computational Chemistry and Materials (CCM)
  - Comp Biology, Pharma, Nanotechnology
- Computational Physics and Electromagnatism
  - National Labs, Defense, Academia Research

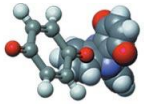
Sample Customer





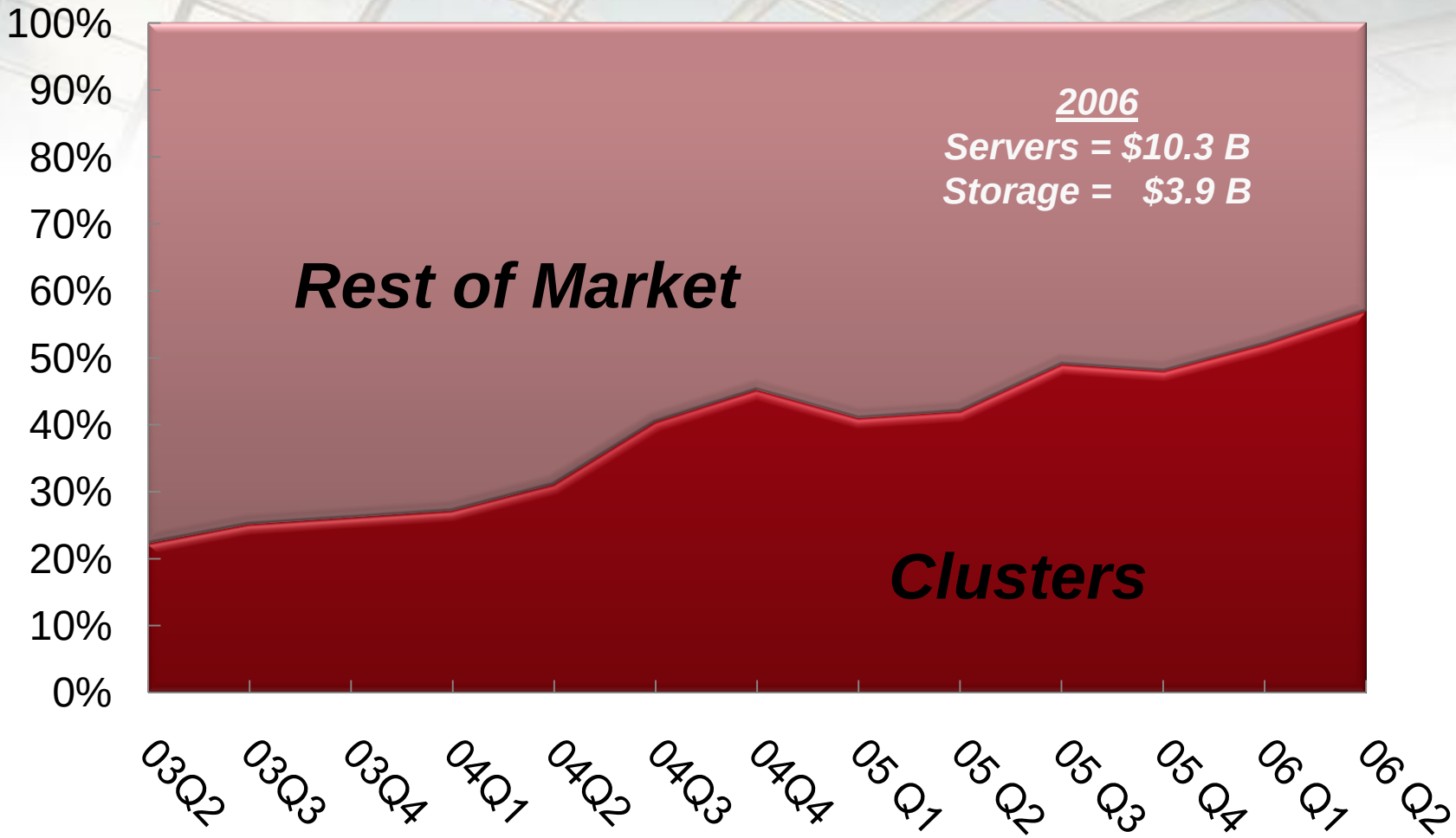






HPC Clusters Drive Relevance of Panasas

IDC: Clusters are now the majority of the HPC market



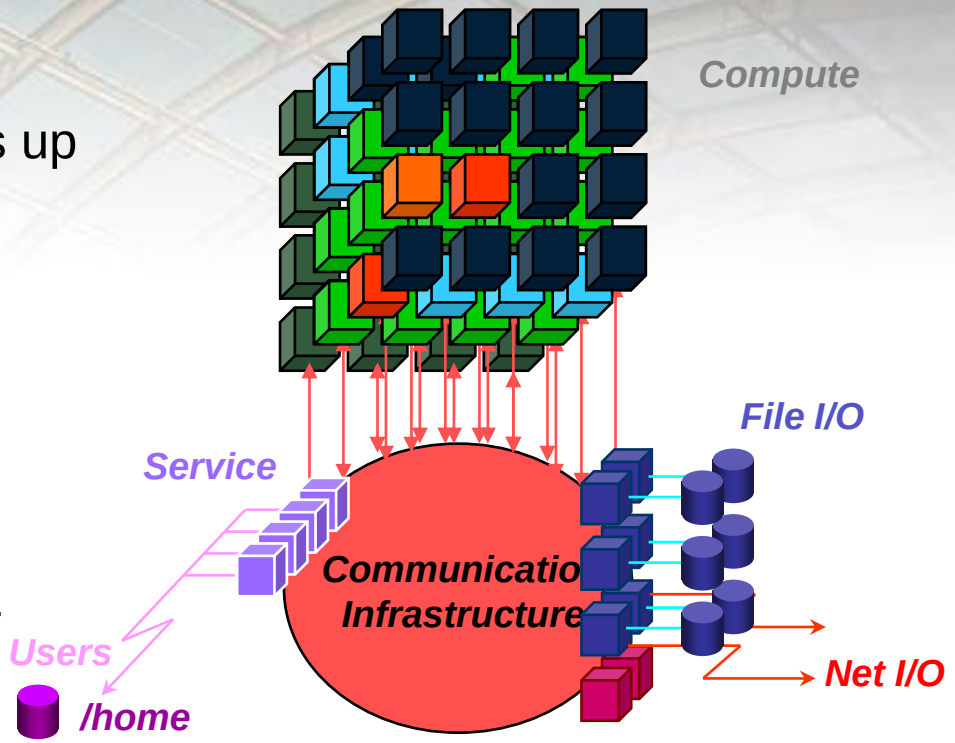
IDC - 2006

Industry Challenges

- Advanced simulation technologies are normally in place and established:
- Generally not in place, but recognised as needs:
  - Repeatable, standardised processes for simulation
  - Ready, structured access to all simulation data, enterprise-wide
  - Seamless integration between “islands” of existing systems and disciplines
  - Systematic means of comparing numerical simulation and physical simulation (test) results

HPC System Challenges

- Setup is painful
  - Takes a long time to get clusters up and running
- Keeping systems updated is difficult
  - Lack of integration into IT infrastructure
- Job management
  - Lack of integration into end-user apps
- Application availability
  - Limited eco-system of application that can exploit parallel processing capabilities



Cluster Computing is now main stream for HPC/Supercomputing

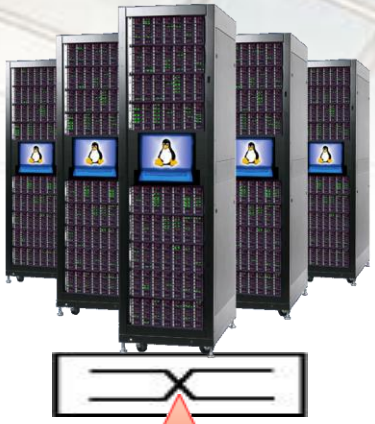
HPC system Requirements

- End customers require:
  - Simple set up and deployment
  - Application availability and integration
  - Simple job submission, status and progress monitoring
  - Compute performance and scalability
- Administrators require:
  - Simplified IT environment
  - Simpler cluster deployment, monitoring, and management
  - Maximum productivity
- Developers require:
  - Maximum productivity programming environment
  - Advanced tools
  - Standards-based environments

How to meet these requirements with Panasas?

Cluster Computing Presents I/O Bottlenecks

Clusters = Parallel Compute



Linux Compute Cluster ***(i.e. MPI apps)***

Single data path to storage

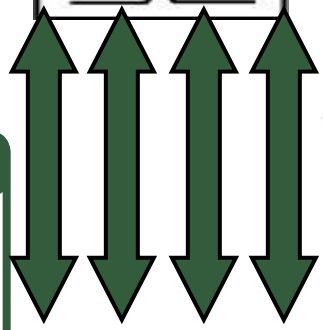
Monolithic Storage ***(NFS servers)***

Issues

- Complex Scaling
- Limited BW & I/O
- Islands of storage
- Inflexible
- Expensive

Parallel Compute needs Parallel IO

Linux Compute Cluster



Parallel data paths

Panasas Storage Clusters

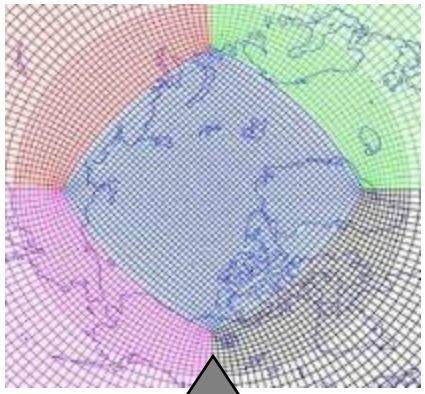
Benefits

- Linear Scaling
- Extreme BW & I/O
- Single storage pool
- Ease of Mgmt
- Lower Cost

I/O Requirement for Supercomputing

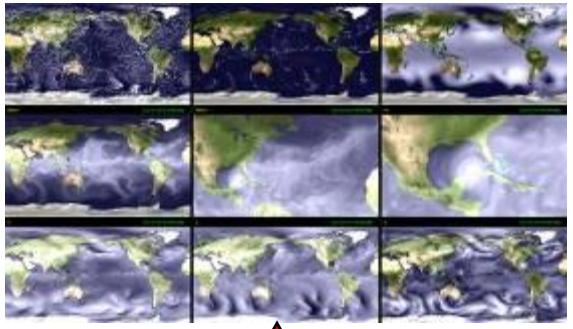
Requirements of Heavy I/O both *Run-Time* and *Interactive* Visualization

Mesh and Models
collaboration



Processing
computation

Visualization
collaboration



Geometry — MB's — OC Database Constraints — GB's — Files — Results — GB's - TB's

Panasas HPC Technology

- Parallel File System
- Unified (Shared) Storage
- Storage and I/O that Scale

(Source MIT DARWIN Project)

Collaboration with Visualization Requires I/O

Panasas and Rocks Visualization Cluster

Collaborators: SU – Panasas – Dell – Cisco – UCSD/Clustercorp

HPCwire
The Leading Source for Global News and Information Covering the Ecosystem of High Productivity Computing / November 15, 2006

Features:

Building a Visualization Cluster with Rocks

by Steve Jones, Lee Shunn, and Tim McIntire
Stanford University Co-Founder, Clustercorp (Rocks)

As high performance computing (HPC) becomes a ubiquitous part of the scientific computing landscape, the science of visualizing HPC datasets has become a critical field of its own. One of the hottest solutions can be found in commoditized high performance visualization clusters (HPVC), which are just starting to pop up in data rich environments around the world.

The widespread adoption of HPC has allowed scientists to process massive multivariate datasets that generate so much data on the output end that the subsequent analysis becomes unwieldy, or in some cases, impossible. These datasets have given rise to great advances in the science of visualization, pushing the limits of traditional workstations beyond their capacity. The only way to visualize this type of output in full-resolution is by building an HPVC that uses a multi-CPU, multi-GPU, multi-display solution to achieve resolution and rendering performance well beyond that of single-system solutions.

This article will take you through each step of building a high performance visualization cluster with a tiled wall display using the Rocks cluster distribution (Rocks Cluster Group, University of California, San Diego), SAGE (Cavern Group at the EVL, University Illinois at Chicago), and Rocks Rolls (from Clustercorp) that integrate commercial CFD applications into the cluster. I/O demands are solved by use of the Panasas file system. The datasets used in this article are from a 65,000 processor run on Blue Gene/L (Lawrence Livermore National Laboratory), with a grid containing 34 billion cells. The visualization cluster is built using Dell, Panasas and Cisco hardware.

Collaboration with Visualization Requires I/O

HPC wire

The Leading Source for Global News and Information Covering the Ecosystem of High Productivity Computing / November 15, 2006

Panasas and Rocks Visualization Cluster

Collaborators: SU – Panasas – Dell – Cisco – UCSD/Clustercorp

Location: HPCC at Stanford, managed by Dr. Steve Jones in support of Flow Physics and Computational Engineering Group

STANFORD UNIVERSITY

clustercorp.com

The Rocks Rolls used for the project:

Visualization:

Viz Roll (from EVL and the Rocks Cluster Group)
EnSight DR Roll (from CEI, Roll by Clustercorp)
ParaView Roll (from ParaView.org, Roll by Clustercorp)

Storage:

Panasas Roll (from Panasas, Roll by Clustercorp)

Networking:

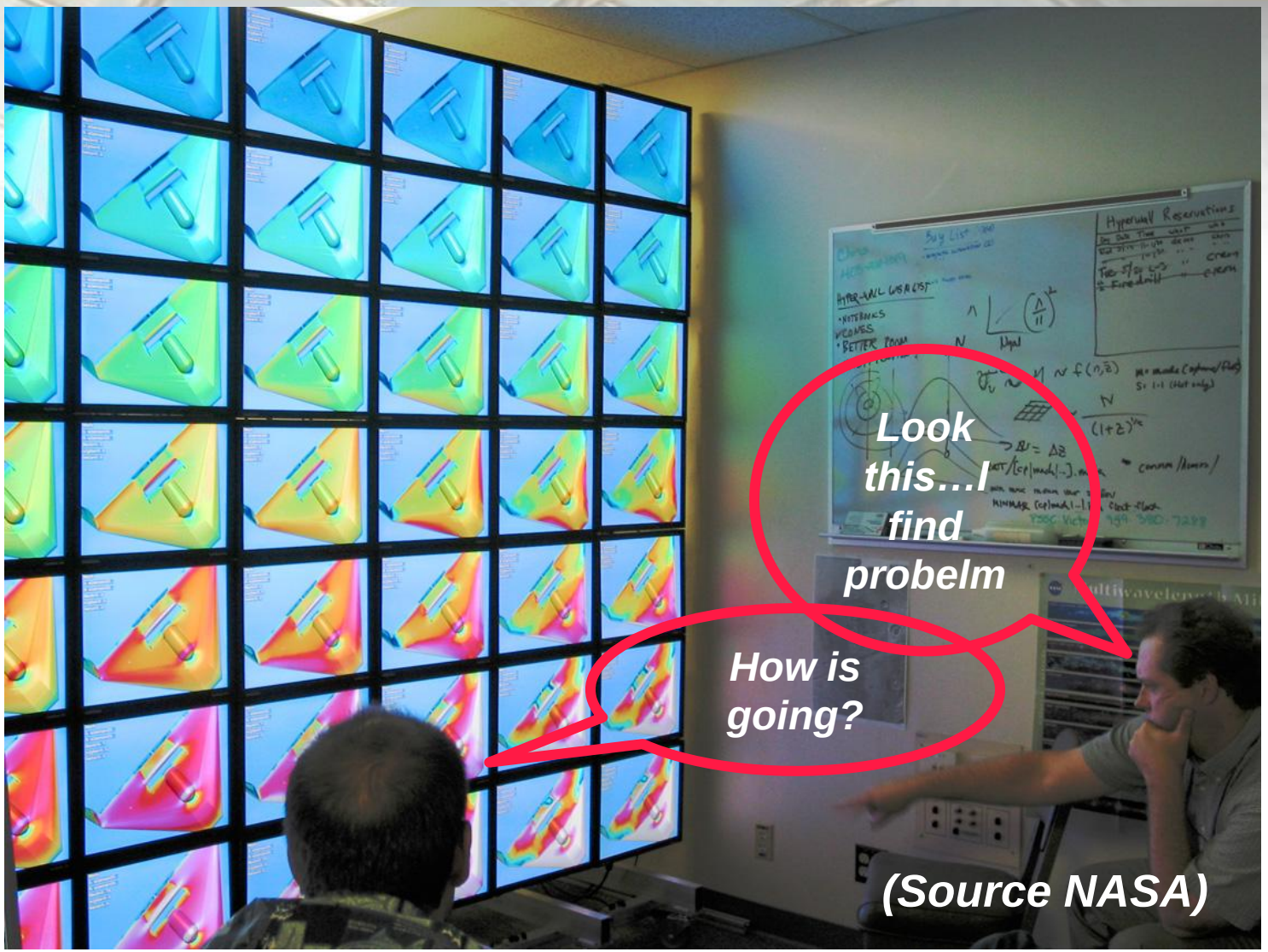
Topspin IB Roll (from Cisco, Roll by Clustercorp)

General:

Kernel, Base, HPC, OS, Web-Server, Ganglia, Java and Service Pack (from the Rocks Cluster Group)
PBS Roll (from the University of Tromso)

Full Article: http://www.hpcwire.com/hpc/1098852.html

parameter studies:Mars Flyer
Columns: increasing mach, Rows:
increasing angle of attack



Example: Productivity Challenges

Simulation Workflow Bottlenecks:

- I/O related to collaboration-intensive tasks:
  - Meshing turn-around requirements for high quality surfaces
  - Post-processing of large files and their network transfer
  - Process integration and automation (model parameterization)
  - Case and data management of simulation results
  - Large dataset for post-visualization

Simulation Workload Bottlenecks :

- I/O related to compute-intensive tasks:
  - Thru-put of “mixed-fidelity” competing for same I/O resources
  - Transient simulation with increased data-save frequency

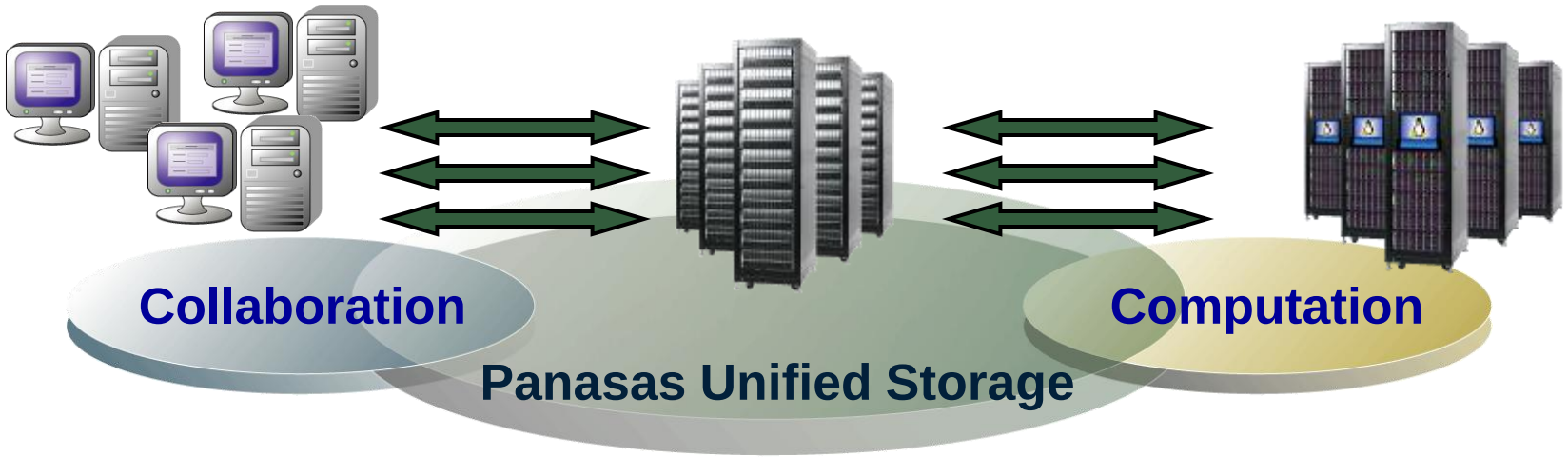
Panasas Offers the Opportunity to Configure I/O and Storage that Targets Specific Simulation Workflow and Workload Challenges

Panasas Response: Unified Storage

- Panasas Unified Storage for an Integrated HPC Workflow
  - Performance for batch run-time I/O, and interactive response and collaboration
  - Management simplicity that enables flexibility in HPC workgroups and workloads
  - Reliability for rapid deployment in exiting HPC production infrastructures

Interactive Requirements
- **Model pre- and post-processing**
- **Simulation process and data management**
- **Application and process integration**

Batch Requirements
- **Parallel jobs with run-time I/O**
- **Multiple job instances**
- **Workload through-put**

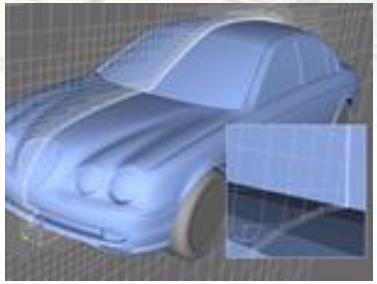


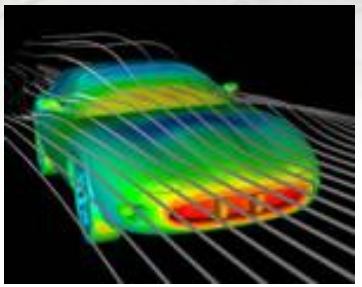
Bottlenecks from I/O in the Workflow

Pre
collaboration

Processing (Solve)
computation

Post
collaboration





MB's

TB's

GB's - TB's

Panasas HPC Technology

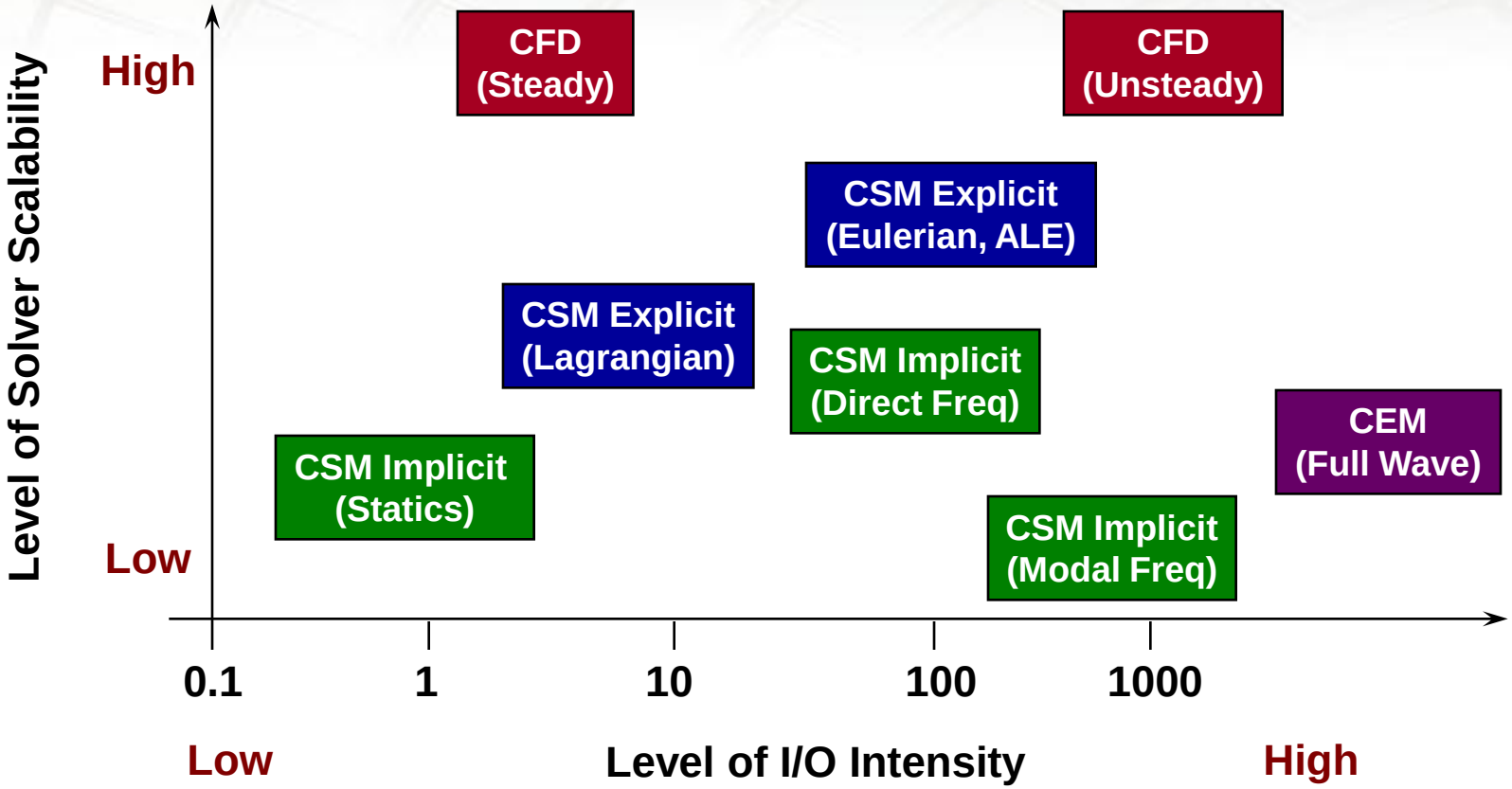
- Parallel File System
- Unified (Shared) Storage
- Storage and I/O that Scale

GB's - TB's

Process and Data Management (PDM) Infrastructure

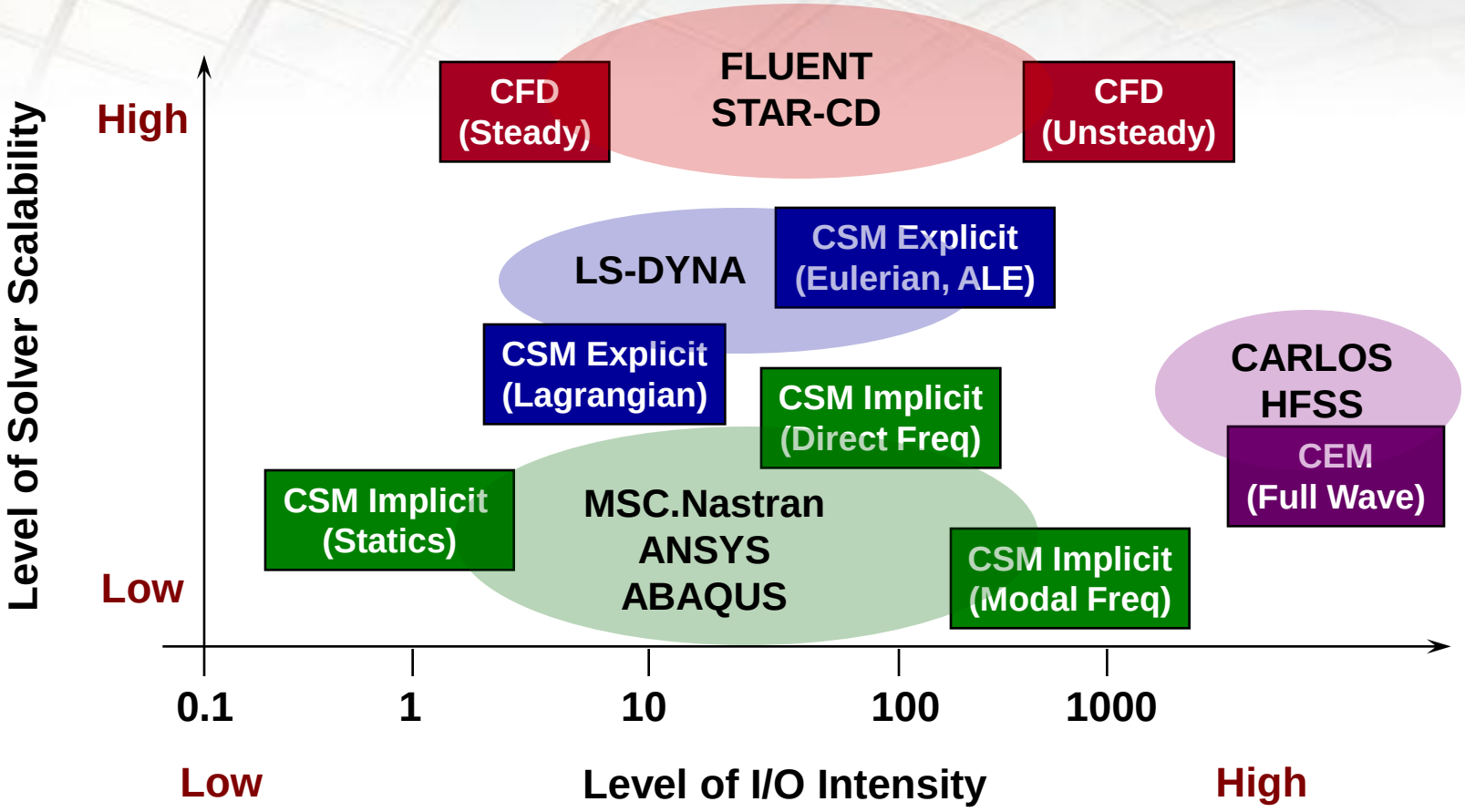
Qualitative Review of HPC Scalability

Typical HPC Behavior of a Single Job by CAE Segment

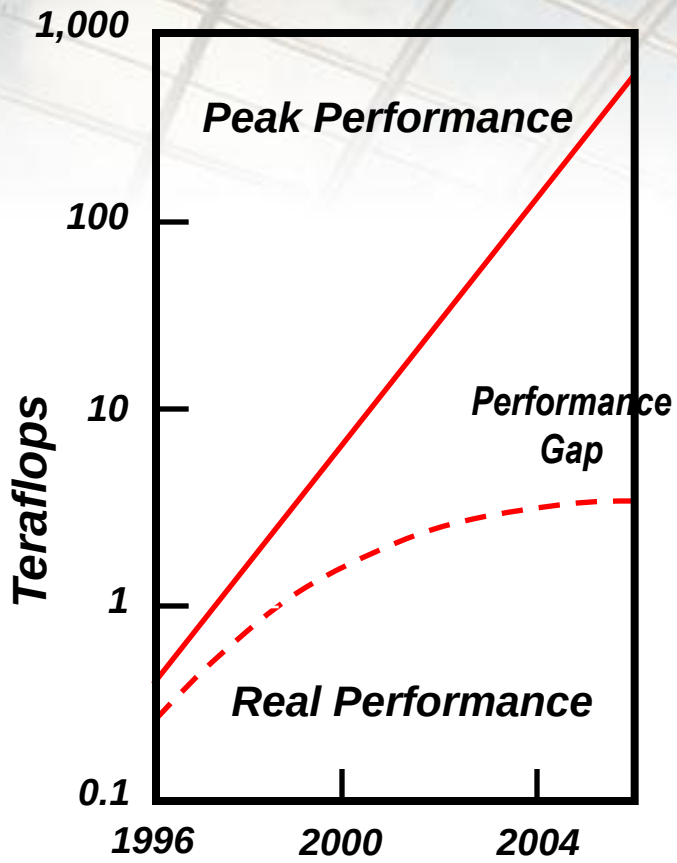


Qualitative Review of HPC Scalability

Typical HPC Behavior of a Single Job by CAE Segment



Peak and real application performance

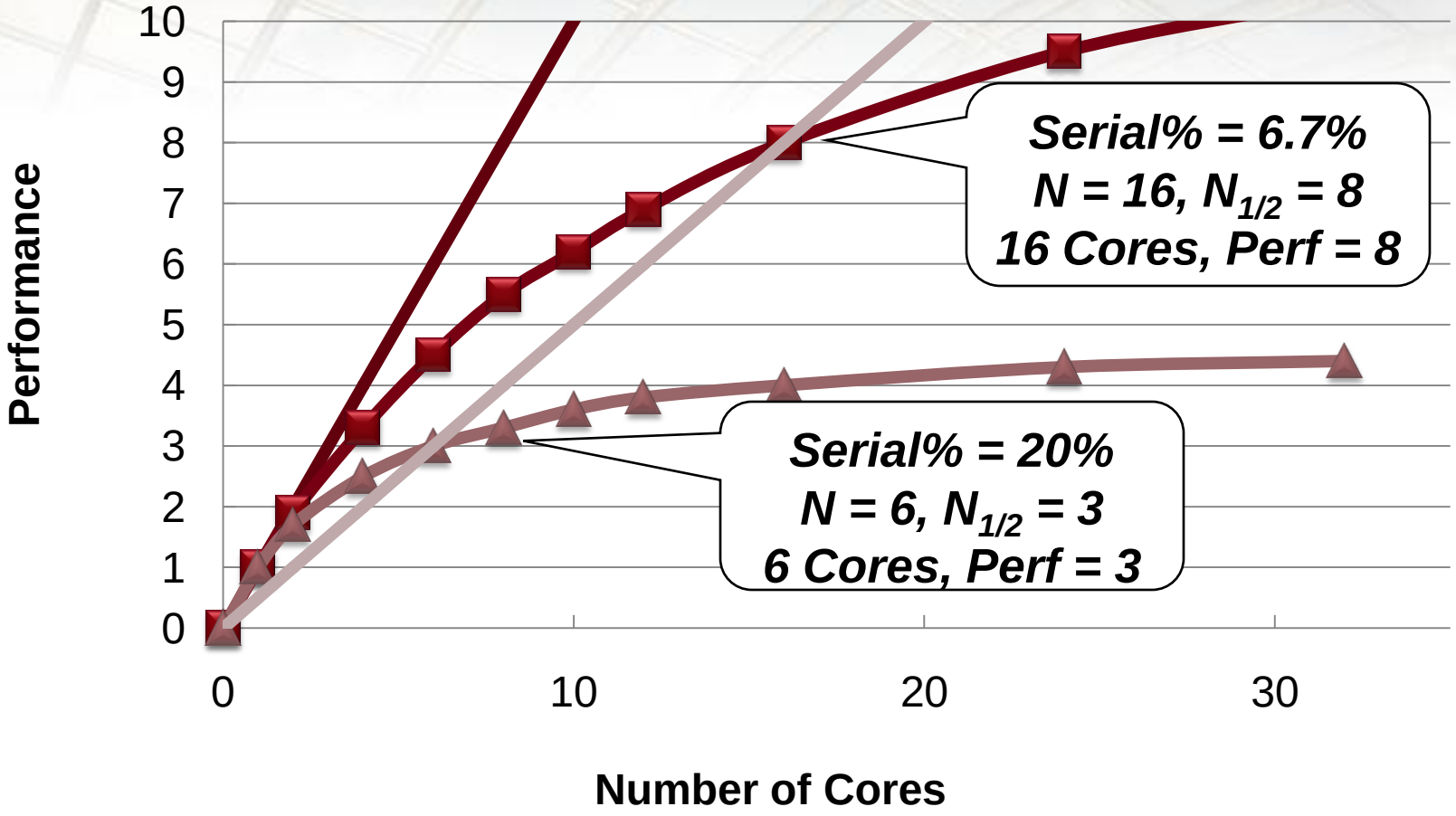


NERSC User Group Meeting June 24-25, 2004
Osni Marques and Tony Drummond
Lawrence Berkeley National Laboratory

- Peak performance improvement
  - 1990 - 2000: 10^2 order
  - 2000 - : 10^3 order
- Sustained performance for HPC applications
  - 1990 – 2000: 40-50% of Peak (Vector Systems)
  - 2000- : 5-10% of Peak
- Solving this performance gap…
  - Higher efficiency algorithm
  - Scalable supercomputing systems including scalable storage and visualization

Application performance scaling

Amdahl's Law: Parallel Speedup = 1/(Serial% + (1-Serial%)/N)

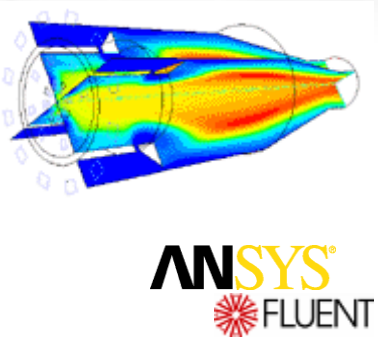


Example: I/O for Transient FLUENT

FLUENT 6.2: FLUENT CFD Transient Case for I/O Study

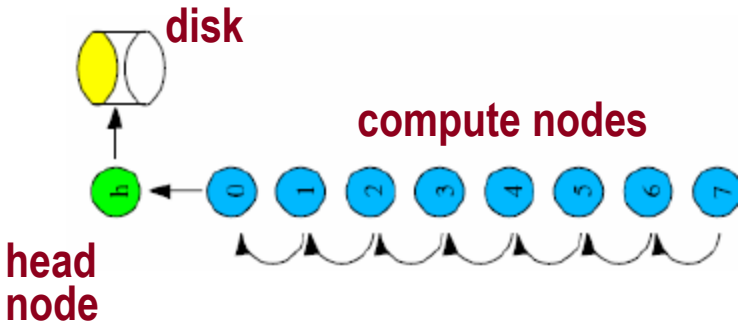
FL5M3 -- Combustion in a High Velocity Burner

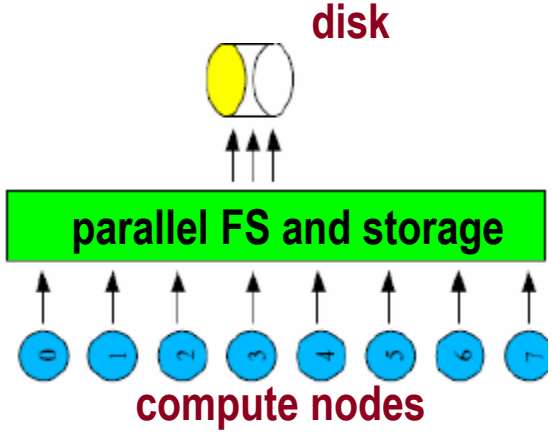
Number of cells	352,800
Cell type	hexahedral
Models	k-epsilon turbulence 6 species with reaction
Solver	segregated implicit



ANSYS FLUENT

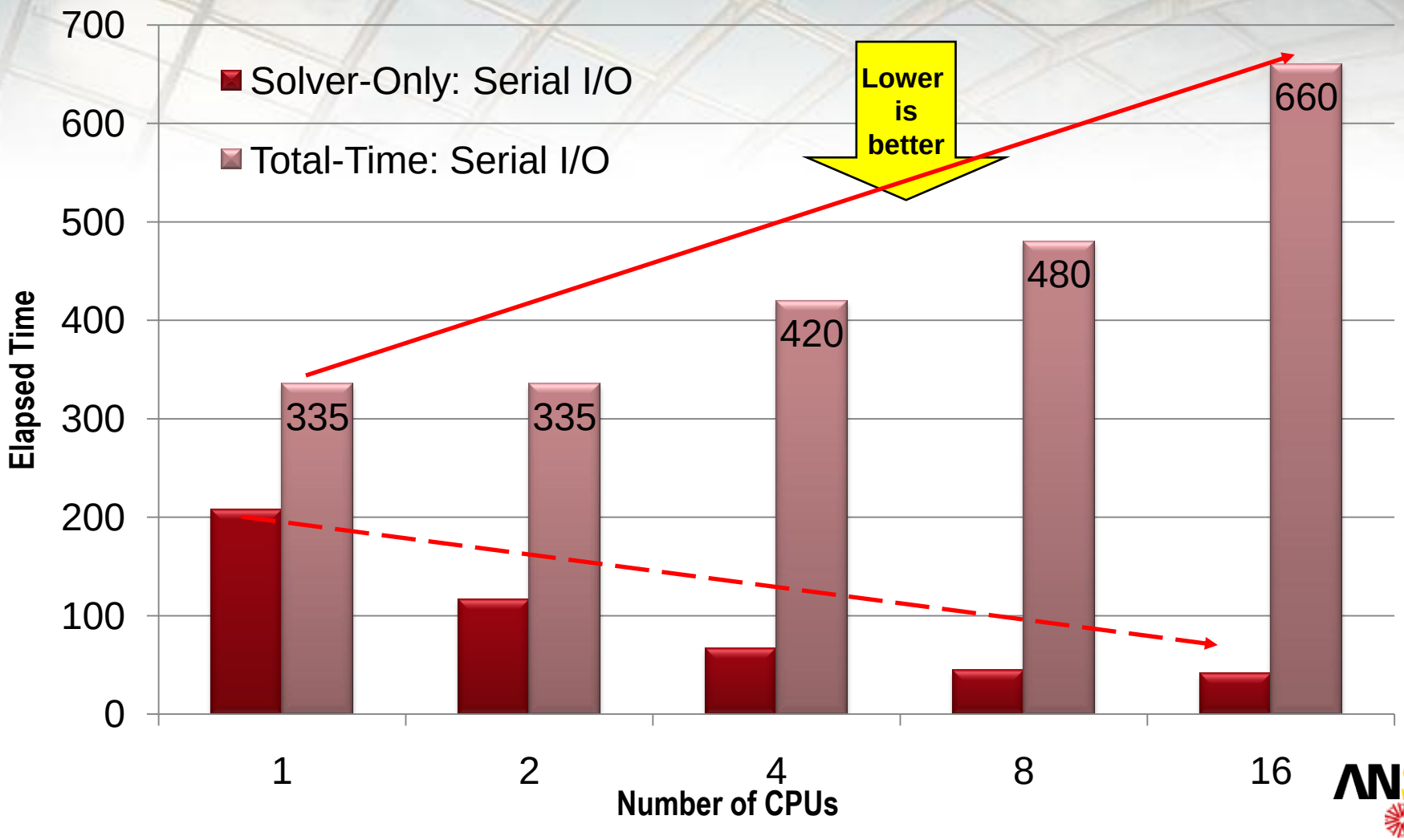
Cluster Configurations





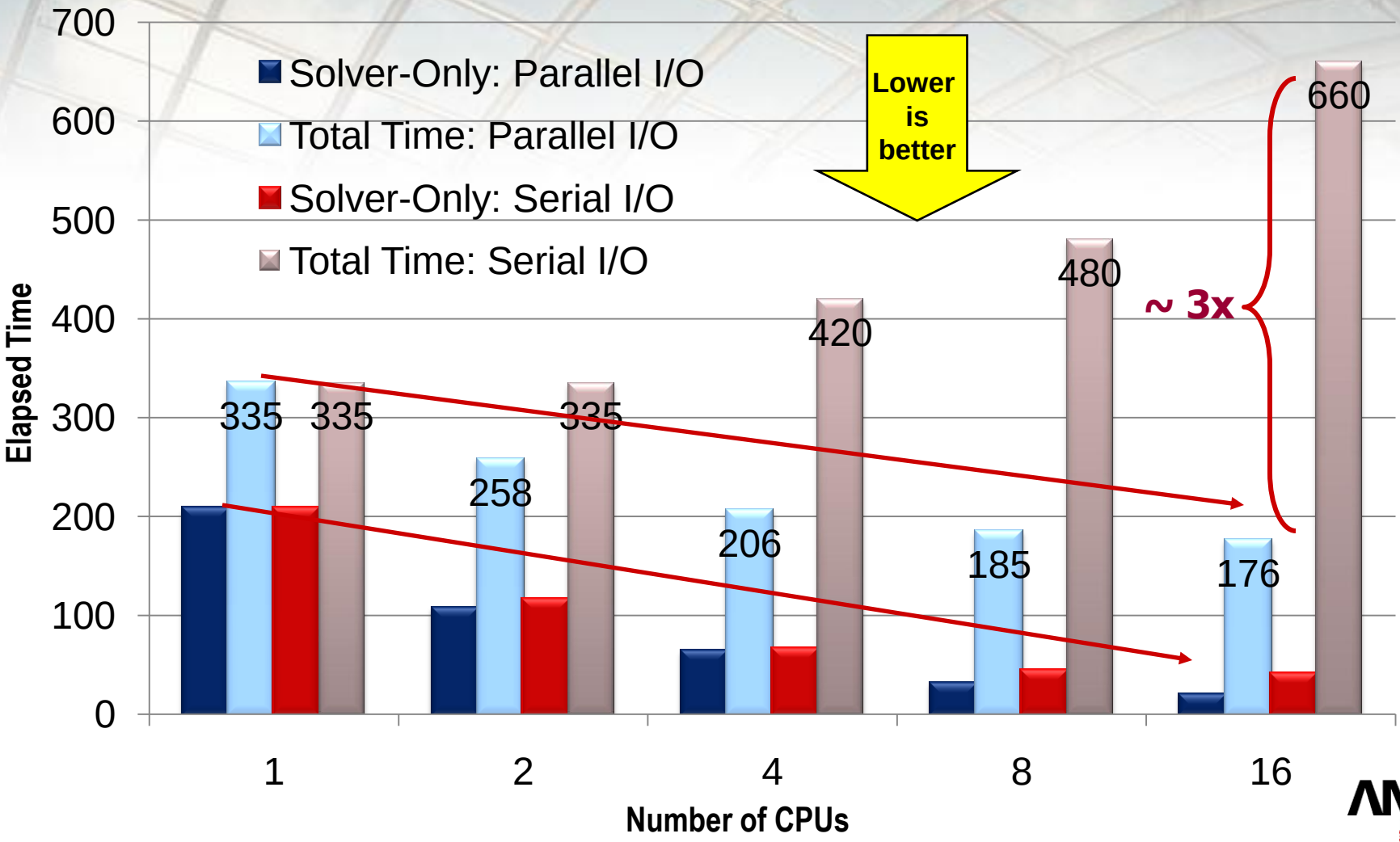
Example: I/O for Transient FLUENT

FLUENT 6.2: FLUENT CFD Transient Case for I/O Study



Example: I/O for Transient FLUENT

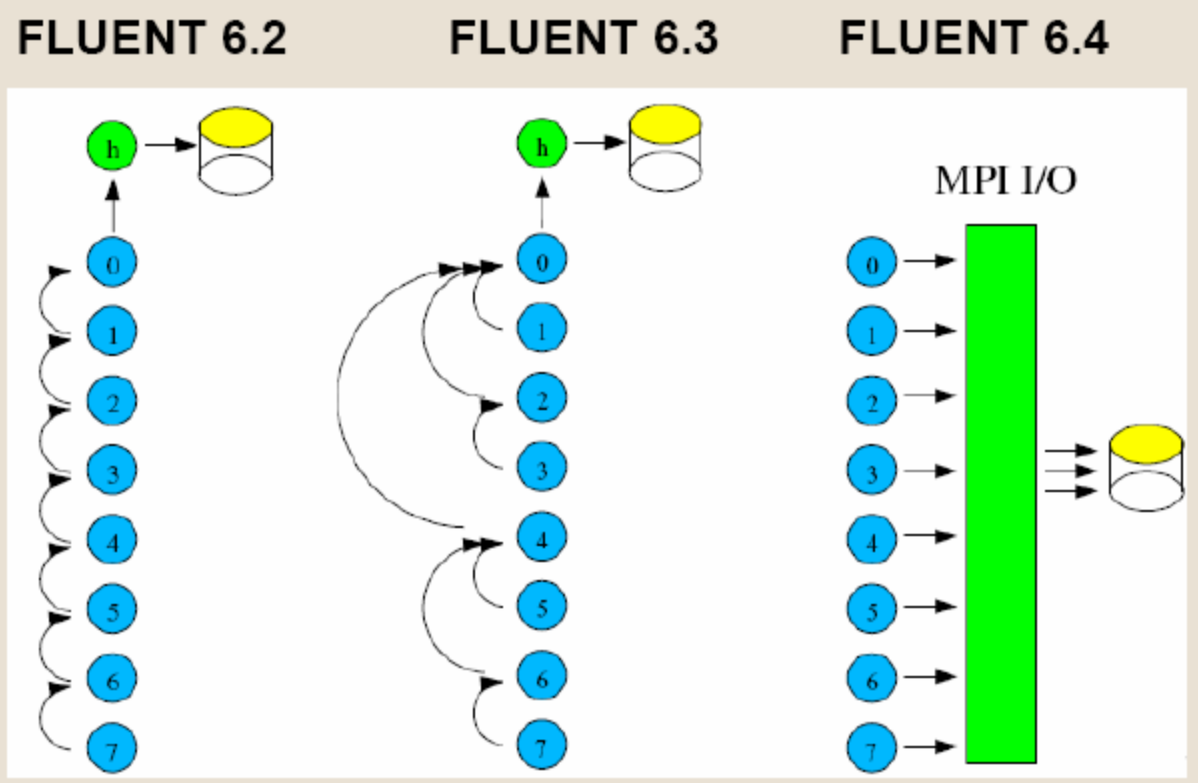
FLUENT 6.2: FLUENT CFD Transient Case for I/O Study



Computation and I/O with FLUENT

Panasas and Fluent Partnership Will Produce Parallel I/O for Future FLUENT 6.4

Serial I/O Scheme **Improvement** **Parallel I/O Scheme**



FLUENT 6.4: (Preliminary) will Support Panasas File System PanFS

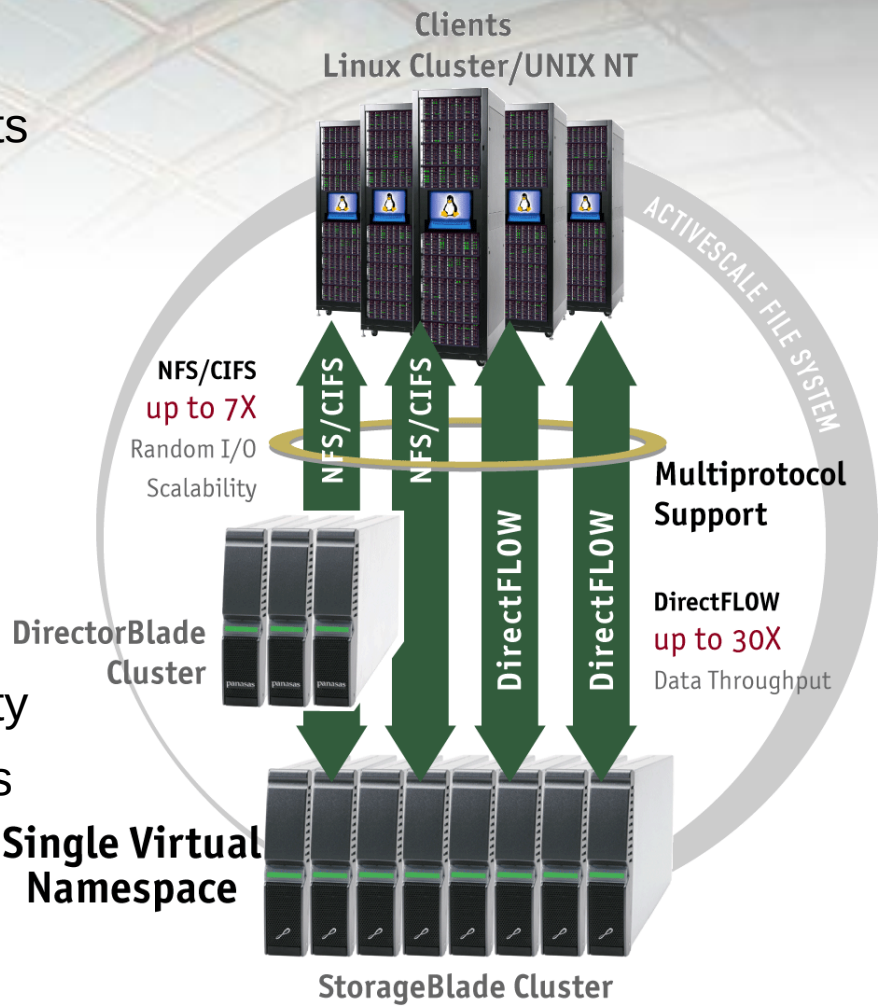
MPI I/O is standard API, but it needs Scalable File System, such as PanFS

Panasas Object-based Storage Architecture

- Object Storage Devices
  - Parallel File system layered over objects
  - Provides scalability, reliability, manageability
- Support for NFS and CIFS

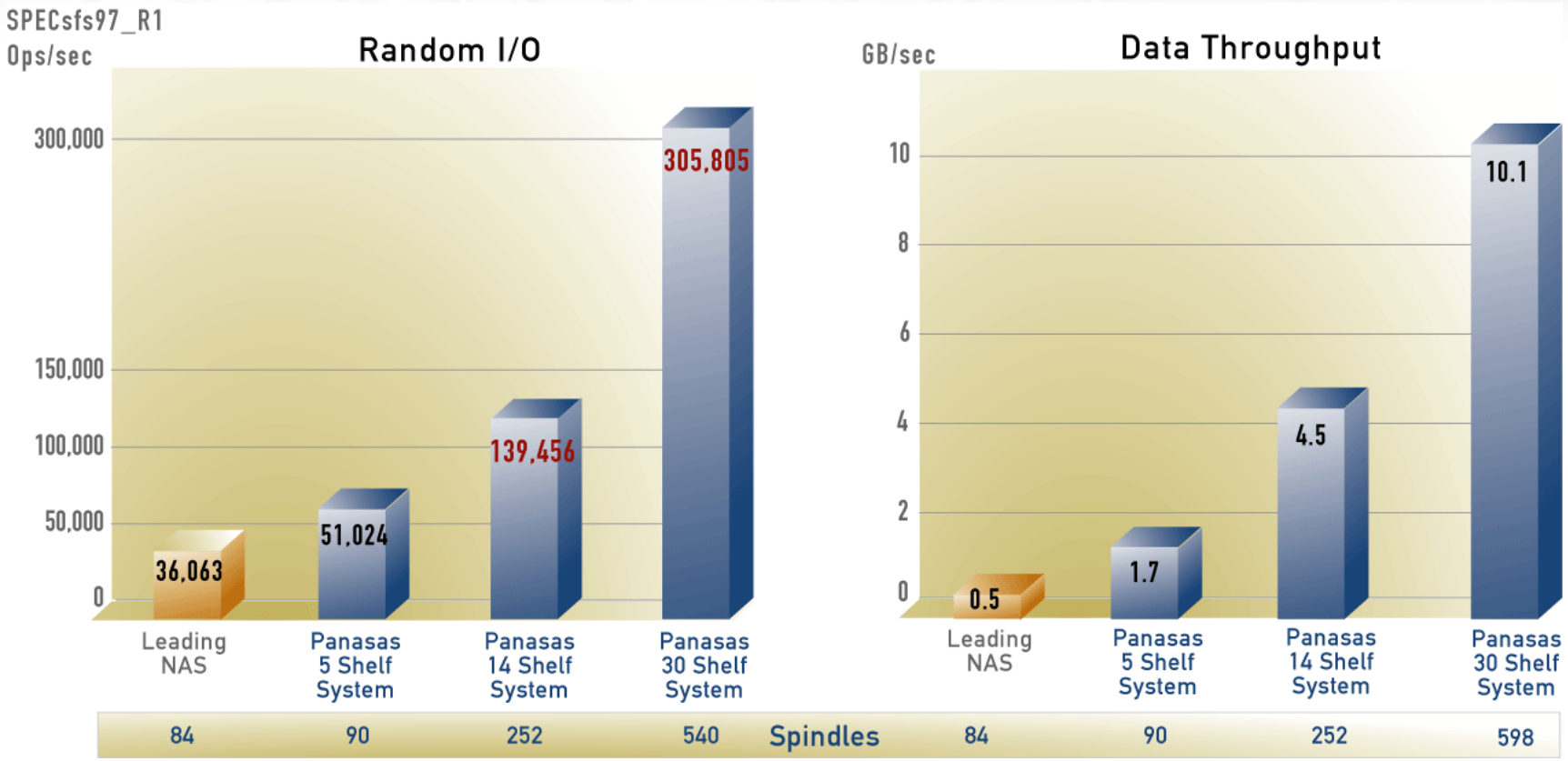
-- OR --

- DirectFLOW client S/W
  - Supports Red Hat, SuSE, Fedora, etc.
- Director Blades
  - Manages & enables metadata scalability
  - Divides namespace into virtual volumes
- Storage Blades
  - Allows wide striping for large files
  - Read ahead/write behind for small files



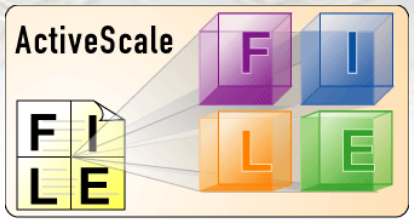
Industry-Leading Performance

- Breakthrough data throughput AND random I/O
  - Tailored offerings deliver performance and scalability for all workloads



Storage Cluster Components

- StorageBlade
  - Processor, Memory, 2 NICs, 2 spindles
  - Object storage system
  - Block management
- DirectorBlade
  - Processor, Memory, 2 NICs
  - Distributed file system
  - File and Object management
  - Cluster management
  - NFS/CIFS re-export
- Integrated hardware/software solution
  - 11 blades per 4U shelf, 5-10 TB/shelf
  - Today: 1 to 30-shelf systems
  - Tomorrow: 1 to 300-shelf systems



Object-based Clustered File System

Smart, Commodity Hardware

Panasas ActiveScale Storage Cluster

Overview of the Panasas ActiveStor

Panasas Storage Cluster Integrates Software and System Solution

Blade-based Storage "Shelf"

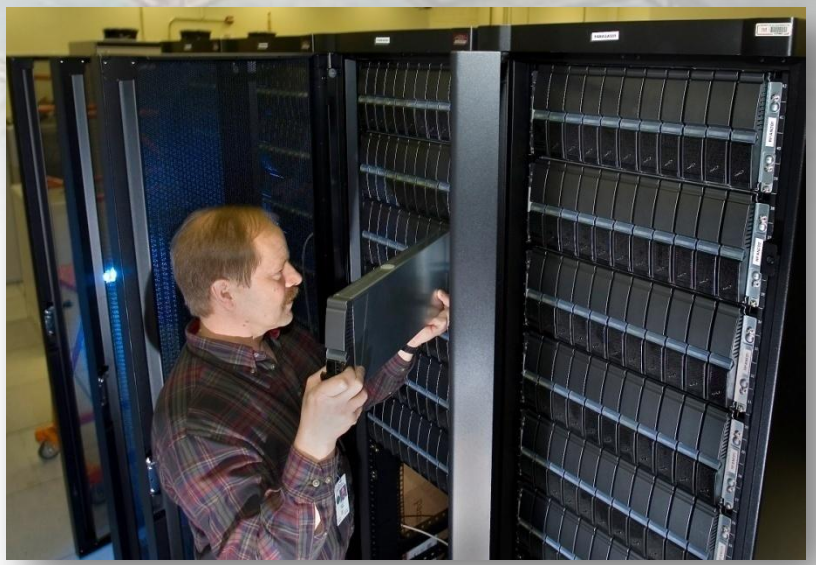
ActiveScale 3.0 Operating Environment

Predictive Self-Management Tools	
DirectFLOW	NFS/CIFS
PanFS	Object RAID

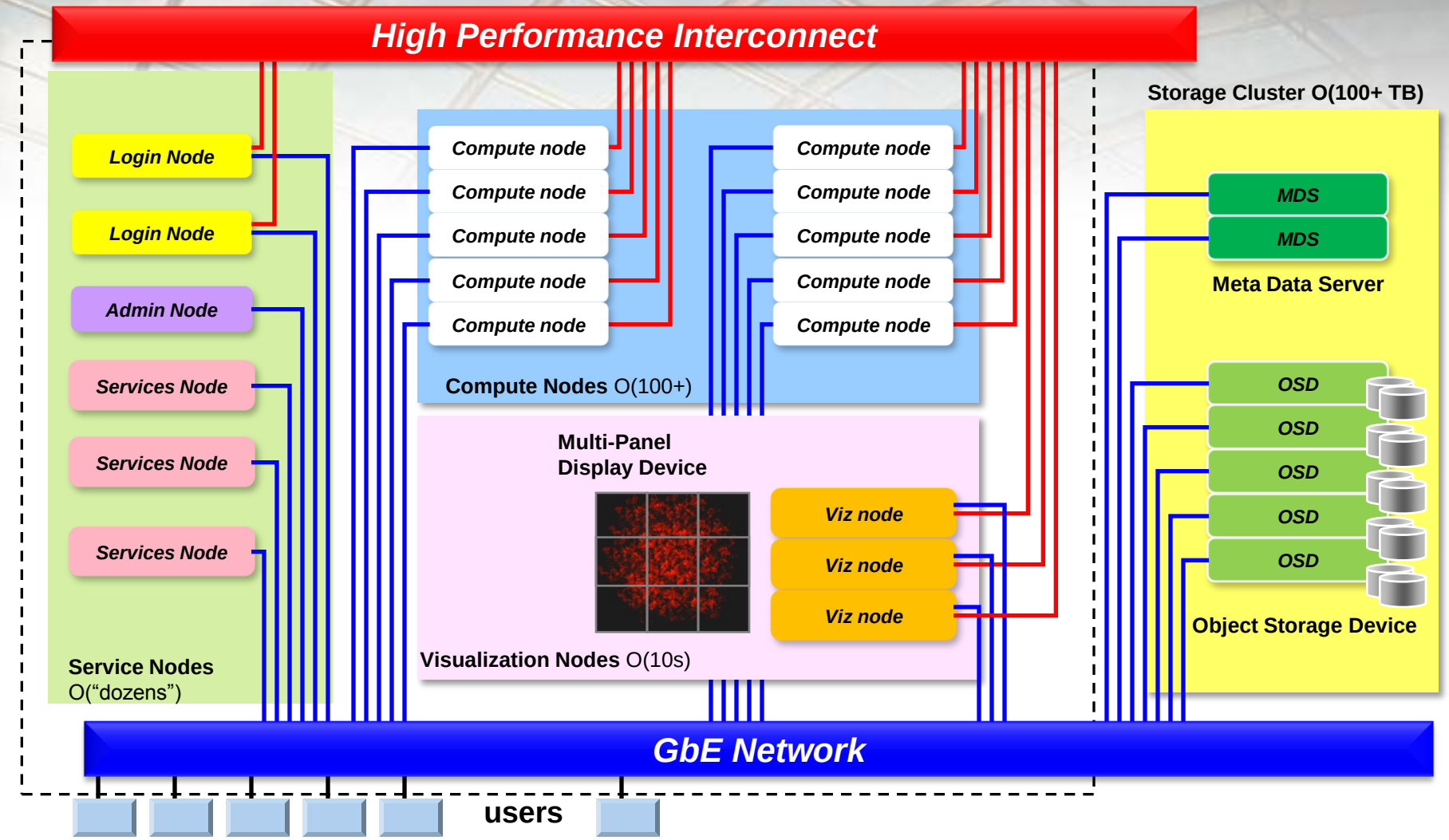
- Features of Single Shelf
  - 11 blades per shelf
  - Up to 1 TB per shelf
  - Up to 20GB cache per shelf
  - Up to 11 StorageBlades
  - Up to 3 DirectorBlades

Panasas Building Blocks

- Objects
  - Container for data and attributes
  - Interface standardized by SNIA T10 as iSCSI/OSD interface
  - Panasas StorageBlade is first commercial OSD in production
- Scalable Panasas RAID
  - Stripe files across container objects
  - Parallel RAID rebuild
- Distributed and Parallel File System
  - Block management hidden behind object storage interface
  - Client IO direct and in parallel to object storage devices
  - File management distributed across metadata managers
  - Robust in the presence of failures



Supercomputing with Panasas



Supercomputing with Panasas

HPC system Requirements

- End customers require:
  - Simple set up and deployment
  - Application availability and integration
  - Simple job submission, status and progress monitoring
  - Compute performance and scalability
- Administrators require:
  - Simplified IT environment
  - Simpler cluster deployment, monitoring, and management
  - Maximum productivity
- Developers require:
  - Maximum productivity programming environment
  - Advanced tools
  - Standards-based environments

Supercomputing with Panasas

- End customers benefit:
  - End user don't need painful 'ftp' for job submission.
  - Sharing large files with compute/visualization/service nodes
  - Windows/Linux client support
- Administrators benefit:
  - Single Unified Namespace
  - Management capability
  - Automatic provisioning for easy growth
- Developers benefit:
  - Scalable standard I/O API
  - Monitoring performance bottleneck for performance improvement.

Supercomputing challenge

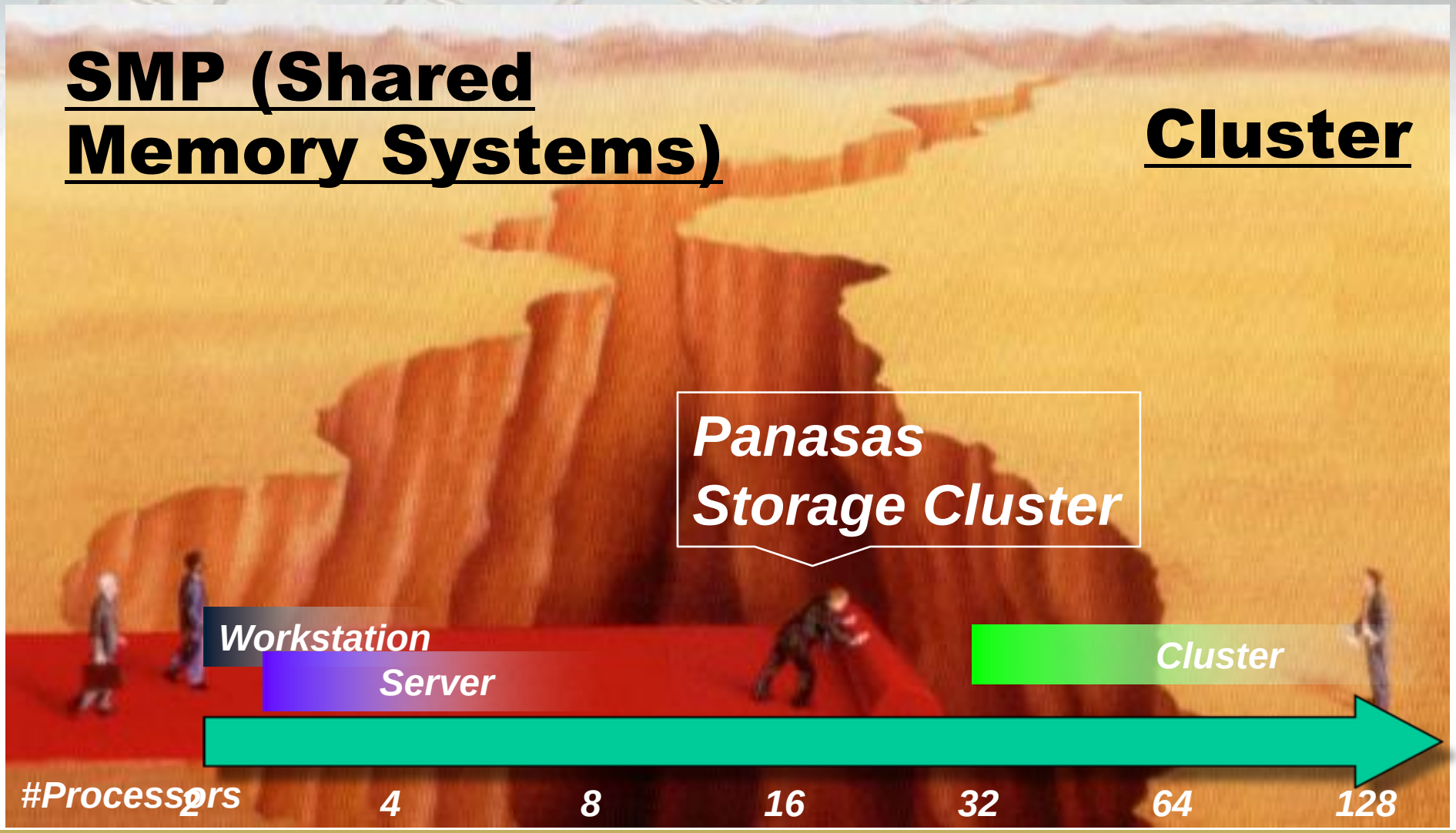
- Current Tera-scale computing problem is not showing real figure of future Peta-scale computing…
- Solving complexity is critical for Peta-scale computing..

Panasas storage cluster can help to solve this complexity problem and support future peta-scale computing challenge.

Panasas Supercomputing Focus and Vision

- Standards-based Core Technologies with Supercomputing Productivity Focus
  - Scalable I/O and storage solutions for computation and collaboration
- Investments in ISV Alliances and HPC Applications Development
  - Joint focus on performance and increased application capabilities
- Established and Growing Industry Influence and Advancement
  - Valued contributions to customers, industry, and research organizations

‘Closing the Gap’



Thank you for this opportunity

Takahiko Tomuro
Technical Consultant for Panasas
Scalable Systems Co., Ltd.
tomuro@sstc.co.jp