panasas

Cluster Scalability of Implicit and Implicit-Explicit LS-DYNA Simulations Using a Parallel File System

Stan Posey
Director, Industry and Applications Development
Panasas, Fremont, CA, USA

Bill Loewe, Ph.D.
Sr. Technical Staff, Applications Engineering
Panasas, Fremont, CA, USA

Paul Calleja, Ph.D.
Director, HPC Services
University of Cambridge, Cambridge, UK

Panasas Overview and LSTC Alliance

- Private, venture-backed company based in Silicon Valley, founded in 1999 by CTO Garth Gibson – a Professor at CMU and Co-inventor of RAID

- Panasas technology combines a parallel file system with a storage hardware architecture for the market's first HPC storage appliance
- Panasas has a global network of direct and channel sales representatives
  - Global resellers include Dell, SGI, and Penguin among others
  - Panasas awarded "Top 5 Vendors to Watch in 2009" at SC08
- Panasas and LSTC have a business and technology alliance since 2006:
  - Panasas has made critical investments in loaner systems and engineering
  - Most leverage with LS-DYNA implicit, but all CAE workloads will benefit
- Panasas and LSTC have many joint customers, samples include:

Select Panasas Customers

panasas

Engineering - Automotive

Engineering - Aerospace

Finance

Life Sciences

Select Panasas Customers

panasas

ConocoPhillips

bp

TOTAL

PGS

PetroChina 中国石油

Energy

3M

intel

SAMSUNG

Pratt & Whitney Canada
A United Technologies Company

Industrial Manufacturing

Walt Disney FEATURE ANIMATION

map SOLUTE

Weta Digital

TRE

Media

Los Alamos
NATIONAL LABORATORY

AWE

NASA

Sandia National Laboratories

Government

CAE Problem Statement: I/O Bottlenecks

panasas

Progression of a Single Job Profile for CAE with serial I/O

1999: Desktops

Typical overnight job – single thread

read = 10 min | compute = 14 hrs | write = 35 min | 5% I/O

CAE Problem Statement: I/O Bottlenecks

panasas

Progression of a Single Job Profile for CAE with serial I/O

1999: Desktops

read = 10 min compute = 14 hrs write = 35 min 5% I/O

NOTE: If we keep read and write times constant in this example . . .

2004: SMP Servers

10 min 1.75 hrs 35 min 30% I/O

CAE Problem Statement: I/O Bottlenecks

panasas

Progression of a Single Job Profile for CAE with serial I/O

1999: Desktops

Typical overnight job – single thread

read = 10 min | compute = 14 hrs | write = 35 min | 5% I/O

NOTE: If we keep read and write times constant in this example . . .

2004: SMP Servers

8-way | I/O significant

10 min | 1.75 hrs | 35 min | 30% I/O

2009: HPC Clusters

32-way | I/O-bound – growing bottleneck

10 min | 26 min | 35 min | *63%* I/O

NOTE: Schematic only, hardware and CAE software advances have been made on non-parallel I/O

Future Use of H/W Accelerators in CAE

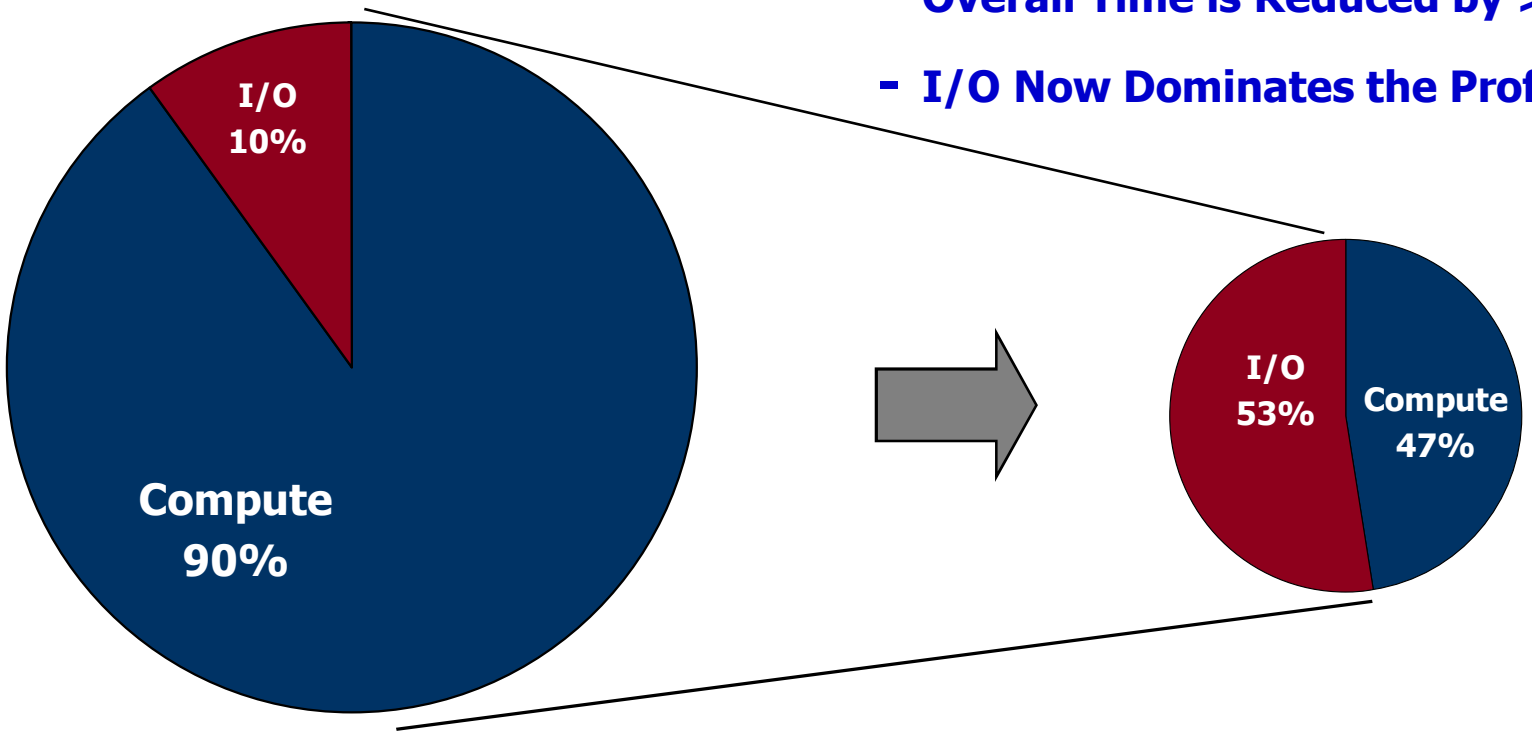
panasas

What Does 10x Acceleration Mean for I/O?

Sample Job Execution Profile of 90% Compute (Numerical Operations) and Modest 10% I/O

What Happens if NumOps are Reduced by Just 10x by Accelerator Technolgy?

- **Overall Time is Reduced by > 5x**
- **I/O Now Dominates the Profile!**

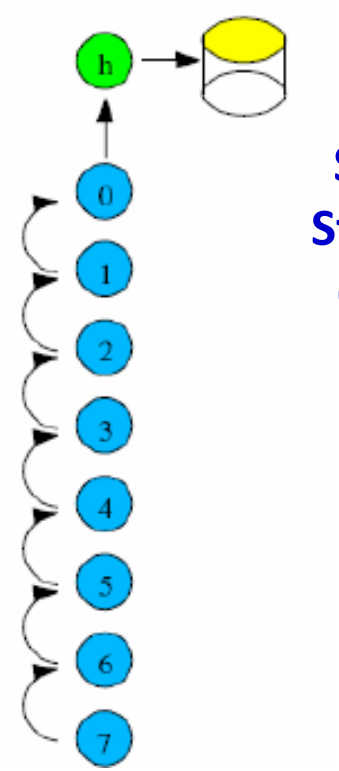


CAE Migrating to Parallel I/O and Storage

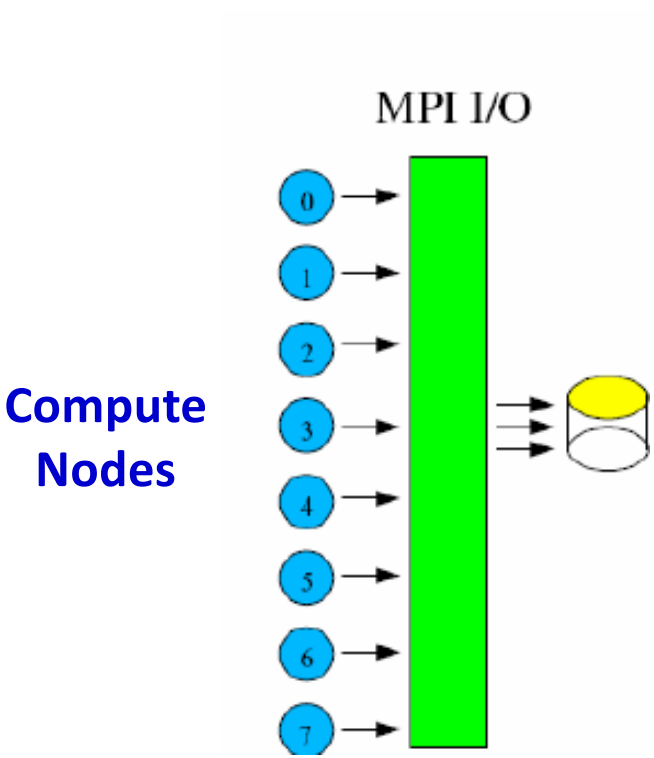
panasas

Schematic of Solution Write for a Parallel CAE Computation

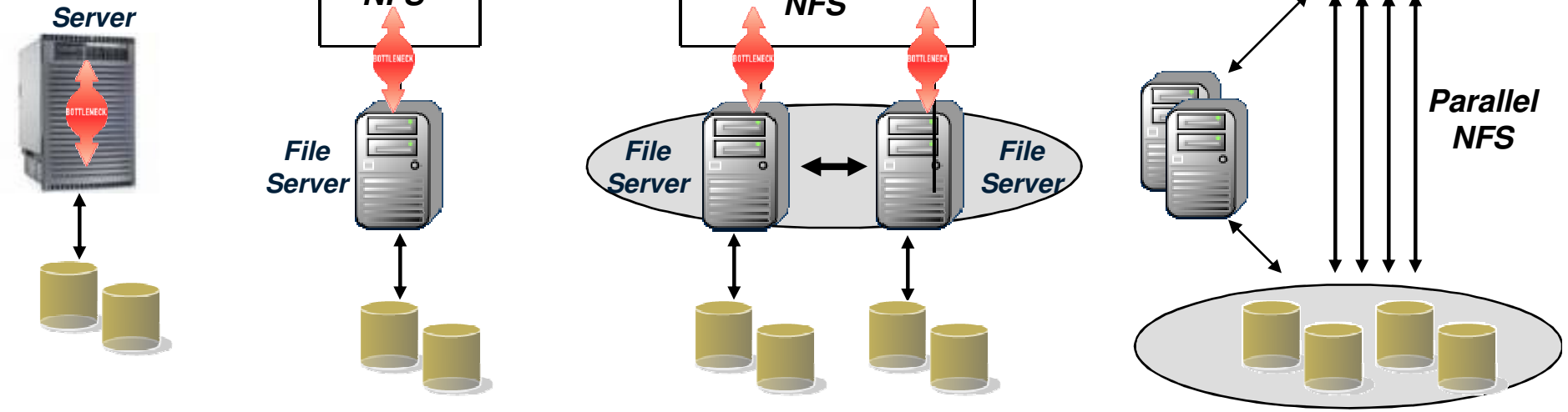
Serial I/O Schematic



Parallel I/O Schematic



Parallel I/O Requires Parallel Storage



DAS: Direct Attached Storage

NAS: Network Attached Storage

Clustered NAS: Multiple NAS file servers managed as one

Parallel Storage: File server not in data path. Performance bottleneck eliminated.

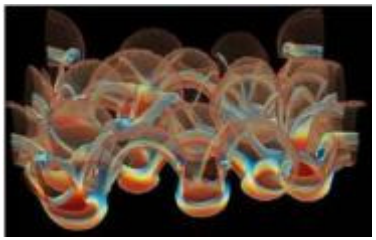
State of HPC: Petaflop Scalability Arrives

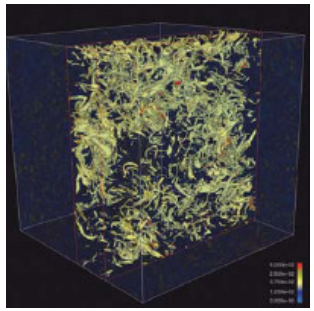
panasas

Los Alamos National Lab

Los Alamos NATIONAL LABORATORY

Advanced Simulation and Computing Center





- **Roadrunner** Tops the Top 500 in June 2008
  - Total 116,640 Cores for DOE weapons research
  - Storage: Panasas PanFS parallel file system and more than 2 Petabytes of capacity

- Applications at Petascale Level
  - MILAGRO – Radiation transport – implicit MC
  - VPIC – Magneto hydrodynamics – particle-in-cell
  - SPaSM – Molecular dynamics of materials
  - Sweep3D – Neutron transport

CAE Solution Focus for Panasas

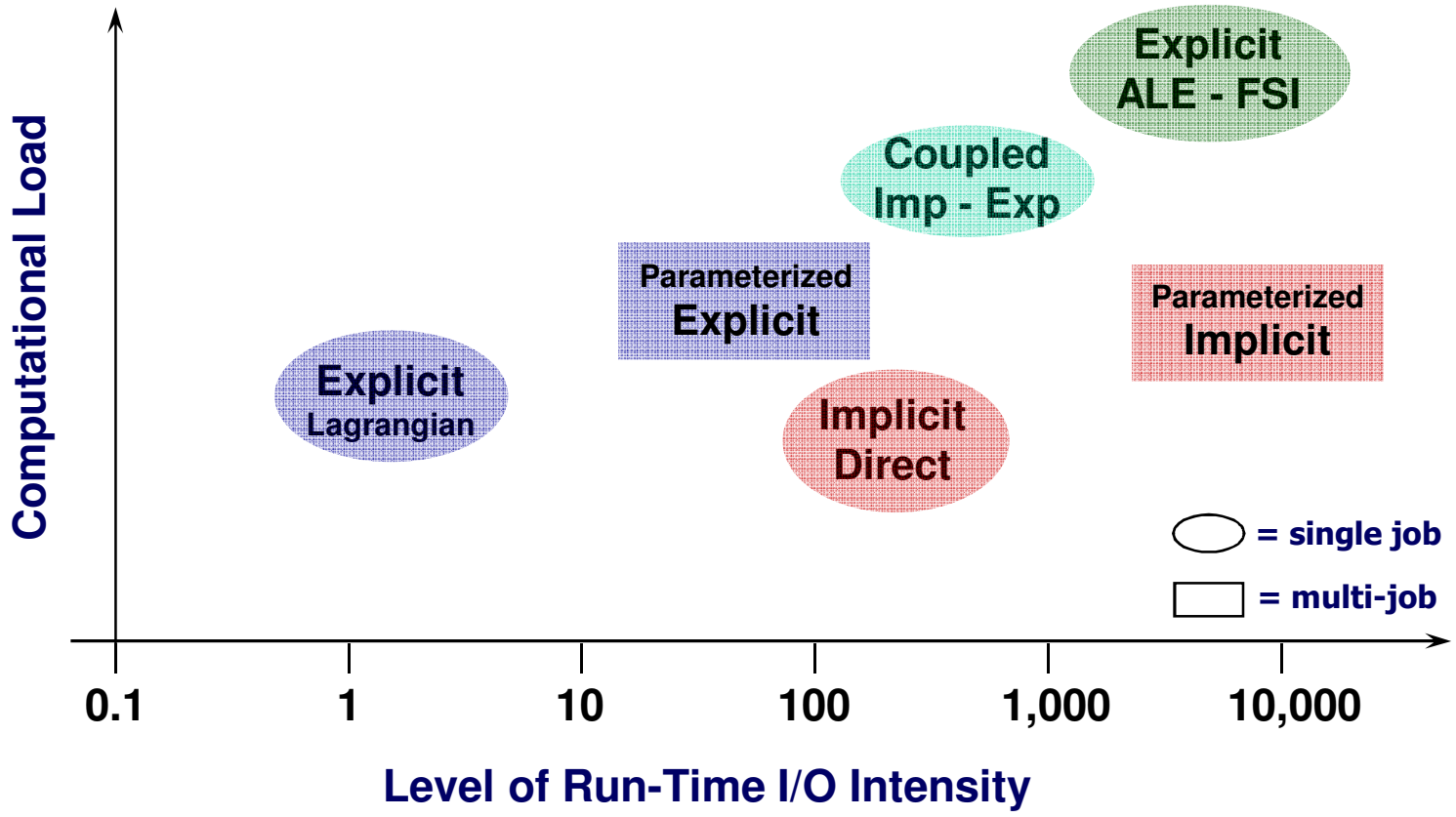
Several CAE Disciplines – Primary Focus is Scalable CFD

- Computational Structural Mechanics (CSM) for Strength; Vibration
  - Strength at minimum weight, low-frequency loading, fatigue
    - ANSYS; ABAQUS/Standard; MSC.Nastran
- Computational Structural Mechanics (CSM) for Impact; Failure
  - Impact over short duration; contacts – crashworthiness
    - LS-DYNA; ABAQUS/Standard; PAM-CRASH; RADIOSS
- Computational Fluid Dynamics (CFD)
  - Aerodynamics; propulsion applications; internal HVAC flows
    - FLUENT; STAR-CD; STAR-CCM+; CFD++; Ansys/CFX; AcuSolve
- Computational Electromagnetics (CEM)
  - EMC for sensors, controls, antennas; low observables RCS
- Process Integration and Design Optimization (PIDO)
  - Simulation environments that couple IFEA, EFEA, CFD, and CEM as required
- CAE Post-Processing and Visualization
  - Qualitative and quantitative interpretation of CAE simulation results

HPC Requirements for CAE Multiphysics

panasas

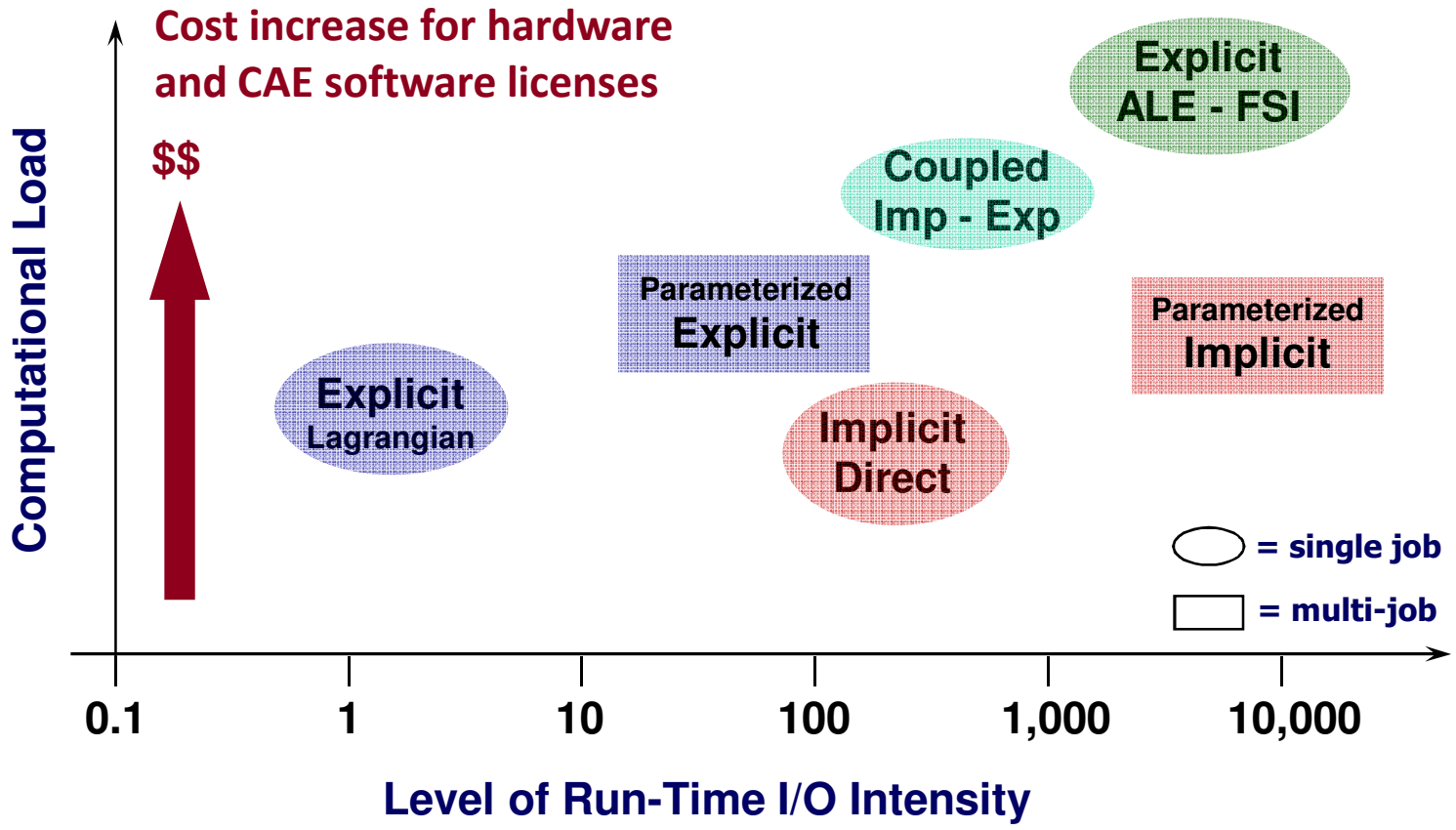
CAE Multiphysics Requires Increasing Computational Load *and* I/O



HPC Requirements for CAE Multiphysics

panasas

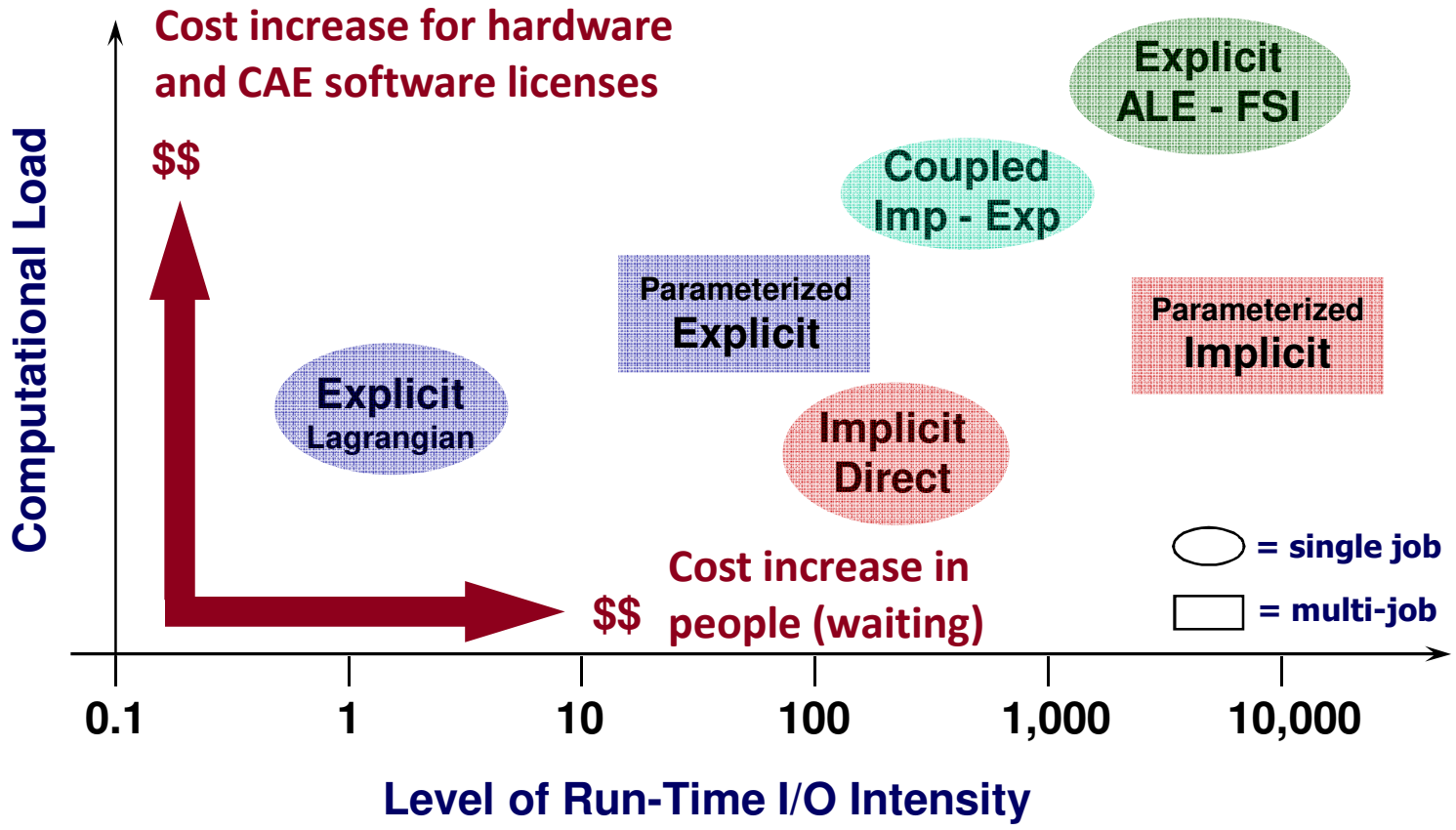
CAE Multiphysics Requires Increasing Computational Load *and* I/O



HPC Requirements for CAE Multiphysics

panasas

CAE Multiphysics Requires Increasing Computational Load *and* I/O



CAE Data-Intensity Challenges Growing

panasas

CAE *Workflow* Bottlenecks:

I/O related to end-user *collaboration-intensive* tasks:

- Long times in pushing sub-domain partitions to nodes
- Post-processing of large files owing to their network transfer
- Case and data management/movement of CAE simulation results

CAE *Workload* Bottlenecks :

I/O related to parallel cluster *compute-intensive* tasks:

- Thru-put of "mixed-disciplines" competing for same I/O resource
- Transient CFD (LES, etc.) with increased data-save frequency
- Large-DOF CSM implicit with out-of-core I/O requirements
- MM-element CSM explicit with 1000's of data-saves
- Non-deterministic modeling automation and parameterization
- General application of multi-scale, multi-discipline, multi-physics

Panasas File System-Based CAE Studies

panasas

ISV Software	Model Size	#Cores	Advantage
FLUENT 12 ANSYS	111M Cells	128	> 2x vs. NAS
STAR-CD 4.06 CD-adapco	17M Cells	256	1.9x vs. NAS
CDP 2.4 Stanford	30M Cells	512	1.8x vs. NAS
Abaqus/Std 6.8-3 SIMULIA	5M DOFs	multi-job	1.4x vs. DAS
ANSYS 11 ANSYS	SP1 Suite	16	"best NAS "
LS-DYNA 971 LSTC	3M DOFs	16	equal to DAS

Significance of Panasas CAE Studies

panasas

- These are commercial applications -- not benchmark kernels
- These studies focus on serial vs. parallel file system benefits
- All CAE models/inputs were relevant to customer practice
- Most were run on production systems at customers, others on OEM (e.g. SGI, Dell) or ISV systems, and no runs at Panasas
- All benchmarks were validated either by an ISV or customer
- Among the all "types" of benchmarks, either CFD, I-FEA, or E-FEA, there was consistency among the numerical results for each type
- These studies have strengthened Panasas relationships with ISVs and boosted ISV and customer confidence in Panasas technical abilities and understanding of industry HPC objectives

Description of System at U of Cambridge

University of Cambridge

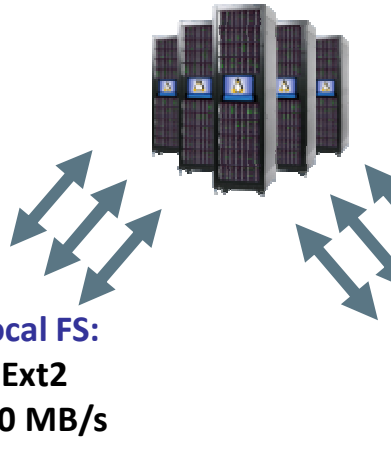
HPC Service, Darwin Supercomputer

Darwin File Systems and Storage

- **PanFS:** 4 Shelves AS3000 XC, 20 TB file system; network connected through Qlogic Silverstorm 9080 and 9240 switches
- **NFS:** Dell PowerEdge 1950 server, Chelsio T310 10Gb ethernet NIC, PERC 5/E RAID, Dell MD 1000 SAS (10TB)
- **Lustre:** v1.6.4.3/DDN storage over Gbit ethernet (87TB)

Univ of Cambridge DARWIN Cluster
Location: University of Cambridge http://www.hpc.cam.ac.uk
Vendor: Dell ; 585 nodes; 2340 cores; 8 GB per node; 4.6 TB total memory
CPU: Intel Xeon (Woodcrest) DC, 3.0 GHz / 4MB L2 cache
Interconnect: InfiniPath QLE7140 SDR HCAs; Silverstorm 9080 and 9240 switches,
File Systems: Panasas PanFS -- 4 shelves AS3000 XC, 20 TB capacity; NFS – Chelsio T310 10Gb ethernet NIC, PERC 5/E RAID Dell MD 1000 SAS 10TB capacity
Operating System: Scientific Linux CERN SLC release 4.6

DARWIN 585 nodes; 2340 cores



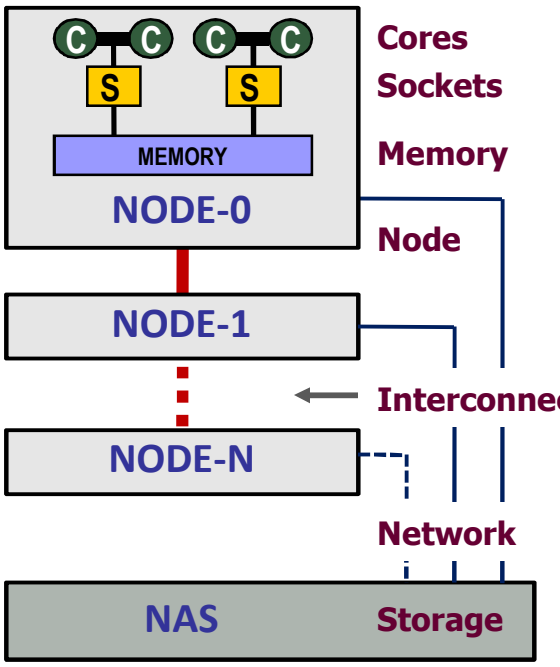
HPC Characterization of an LS-DYNA Job

panasas

Like most all parallel FEA, an LS-DYNA job contains a mix of compute tasks that each require specific performance attributes of an HPC system:

- **Numerical Operations:** typically equations solvers and other modeling calculations
- **Communication Operations:** partition boundary information "passed" between cores
- **Read and Write Operations:** case and data file i/o before/during/after computations

Schematic of HPC System Stack



LS-DYNA Compute Task

- **Numerical Operations**
- **Communications (MPI)**
- **Read/Write Operations**

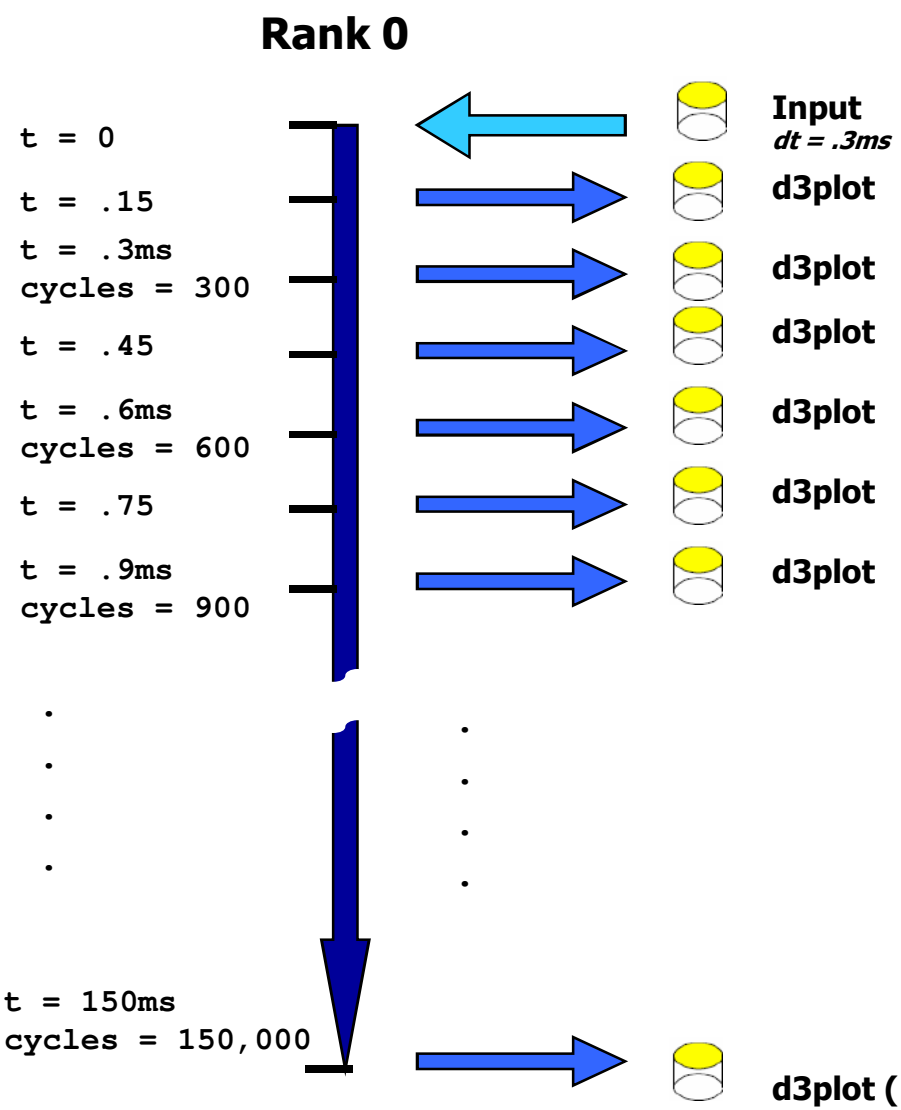
Performance Attribute

Fast CPU architectures to speed-up equation solvers operating on each partition

Low-latency interconnects and MPI system software to minimize communications overhead between partitions for higher levels of scalability

Parallel file system with NAS to ensure concurrent reads and writes that scale the I/O

Increasing I/O in LS-DYNA User Practice



LS-DYNA Higher Output Frequency and Data Size

- today limited by I/O bottlenecks to MPI rank 0
- desire for improved user understanding of the event evolution
- desire to monitor solution for error (contact issues, element distortions, etc.)

3car Performance for 8 Cores x 2 Nodes

panasas

LS-DYNA 971: Comparison of PanFS vs. Local FS

LSTC

panasas

PanFS -- Total Time

Local FS -- Total Time

Singe Job on Single 16-Core Node

Total Time in Seconds

14400

10800

7200

3600

0

Lower is better

11640

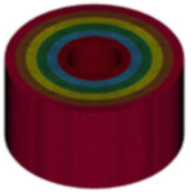
11768

700K Elements 3 Car Model

NOTE: PanFS Advantage in Total Time vs. the Local FS

16 - Cores

LS-DYNA Implicit I/O Scheme



Job Task	IO Scheme	IO Operation
start	Input -- serial IO	Read nodes, elements and control file
element matrix generation and assembly into global matrix		
matrix factor + fsb (dominant phase, Can be as much as 90% of total time)	Scratch -- parallel IO	Factor matrix out-of-core, reads/writes
. . .	. . .	
stress recovery, multiple RHS's		
complete	Results -- serial IO	Write solution results [100's of GB's of I/O]

CSM implicit solver LS-DYNA is direct and single-step, with out-of-core READS and WRITES

-- I/O occurs in the sparse factor phase of the solver

-- this scheme is for static, if an eigensolution (Lanzcos), then I/O can be VERY heavy

Performance for LS-DYNA 971 Implicit

panasas

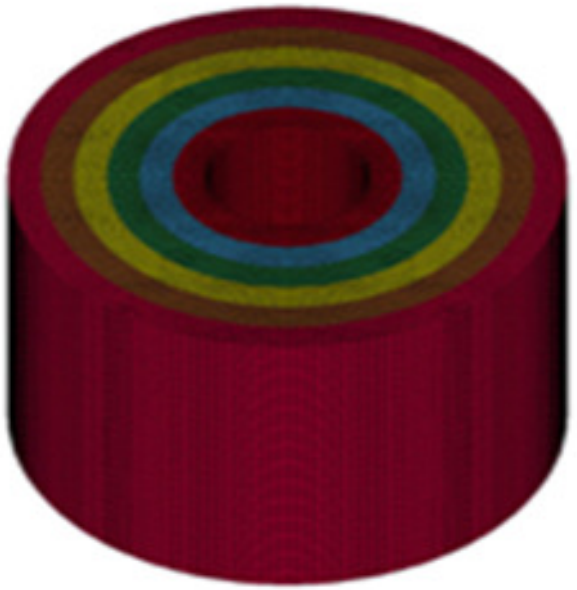
LS-DYNA 971: Comparison of PanFS vs. Local FS

LSTC

panasas

Benchmark Problem – CYL1E6

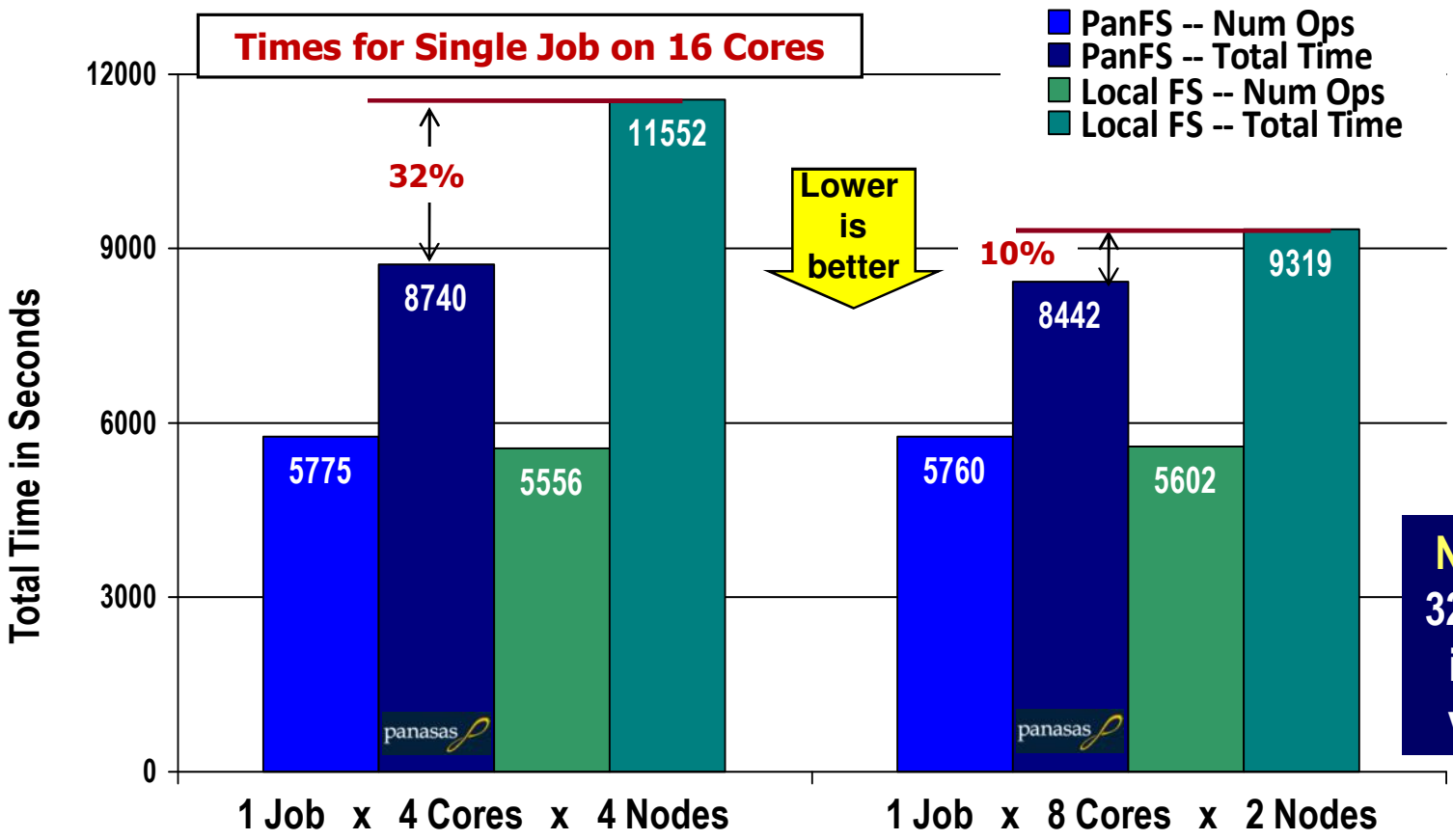
- LS-DYNA v971 implicit
- 6 nested cylinders with contact between them
- 921,600 Solid Elements
- 1,014,751 Nodes
- 3,034,944 Order of Linear Algebra Problem
- 1 Nonlinear Implicit Time Step, 2 Factors, 2 Solves, 4 Force Computations



Performance for LS-DYNA 971 Implicit

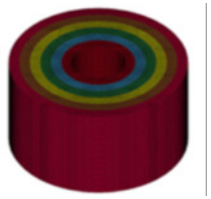
panasas

LS-DYNA 971: Comparison of PanFS vs. Local FS



LSTC

panasas



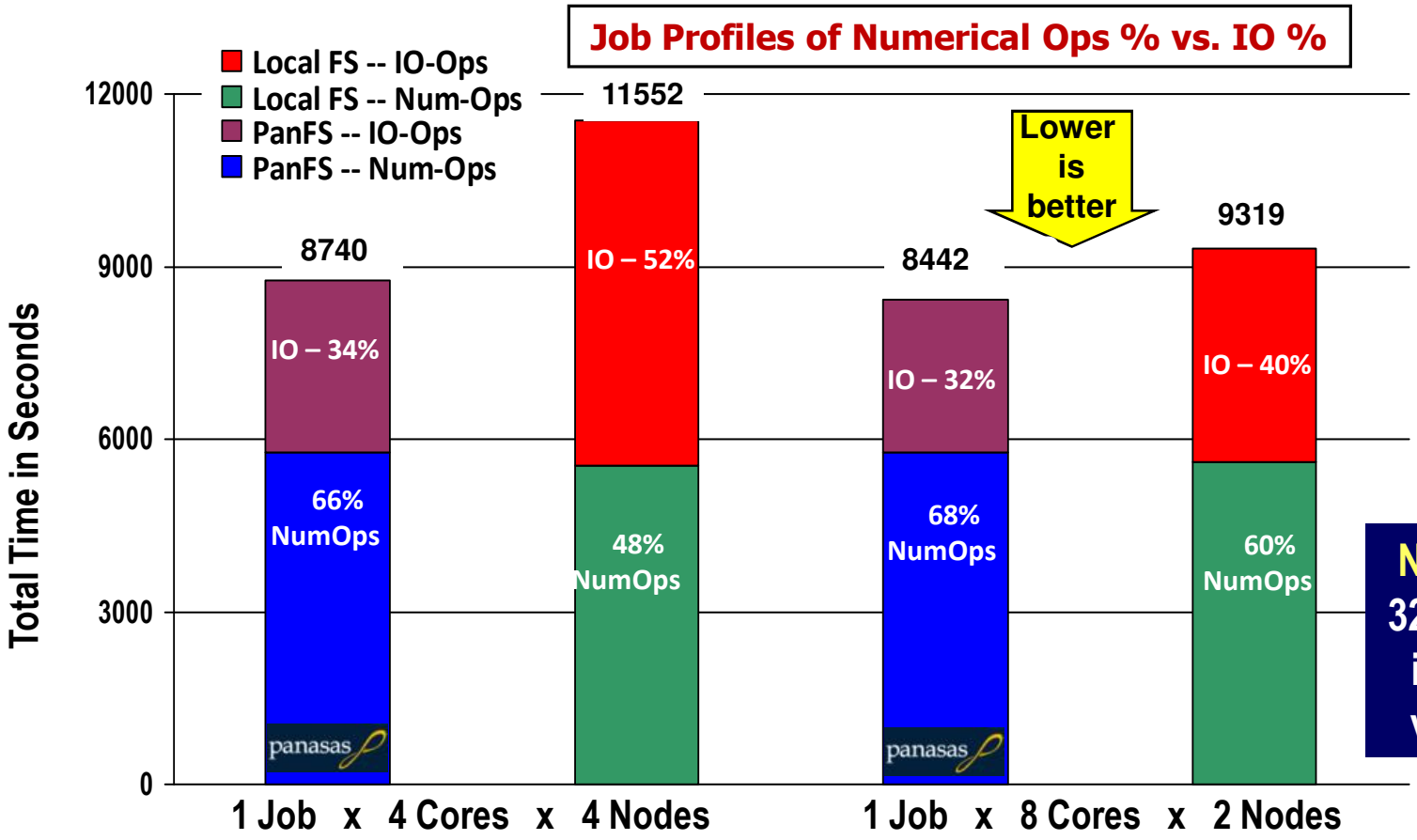
3M DOF Cylinders

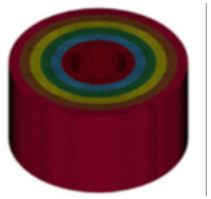
NOTE: PanFS 32% Advantage in Total Time vs. Local FS

Numerical vs. IO Computational Profile

panasas

LS-DYNA 971: Comparison of PanFS vs. Local FS





3M DOF Cylinders

NOTE: PanFS 32% Advantage in Total Time vs. Local FS

Acknowledged Contributors to the Study

University of Cambridge

- Dr. Paul Calleja, Director, HPCS
- Dr. Stuart Rankin, Lead System Manager, HPCS

LSTC

- Dr. Jason Wang, Parallel Development Lead Explicit
- Dr. Roger Grimes, Parallel Development Lead Implicit

Panasas

- Mr. Derek Burke, Director of Marketing, Panasas EMEA

Why Organizations Choose Panasas

panasas

Existing demand for a parallel file-system

- I/O intensive jobs and/or multiple-jobs performing I/O simultaneously and/or a high aggregate I/O bandwidth required

Requirement for a "production-ready" solution

- **Easy to Install**: 1.5 hours to install, configure, and begin running jobs
- **Easy to Scale**: Scales performance with capacity. e.g. 1 shelf provided 600 MB/s; 2 shelves provided 1.2 GB/s. Dynamically load-balances data as additional capacity is added without disruption

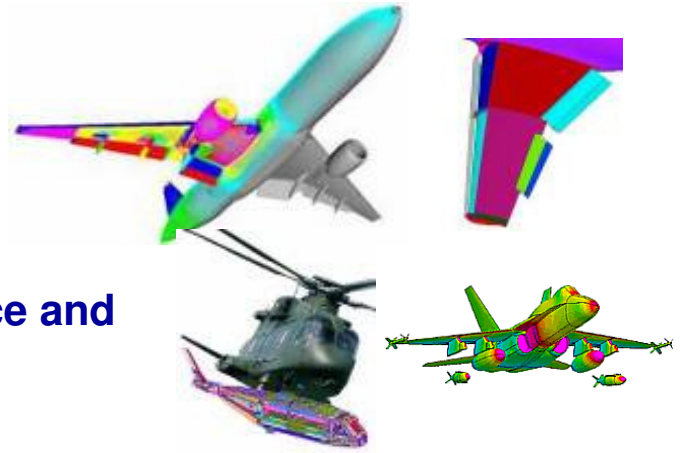
Very competitive total cost of ownership

- **Best Value:** For less than the fully burdened cost of NAS
- storage, you can get HPC storage from Panasas
- **Easy to Manage**: Extremely easy to administer; very low
- administration costs

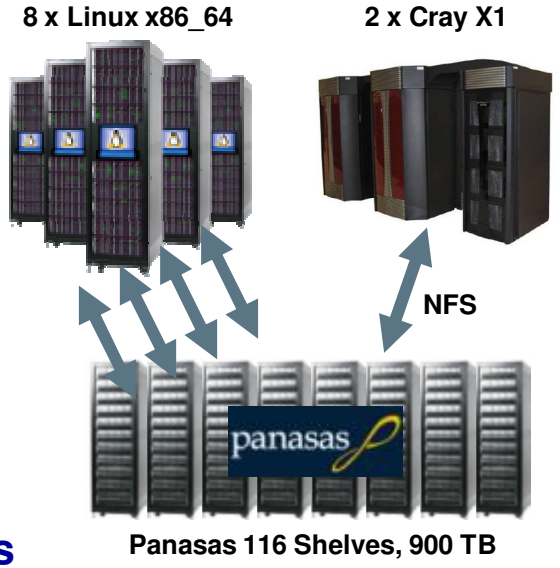
Boeing HPC Based on Panasas Storage

Boeing Company

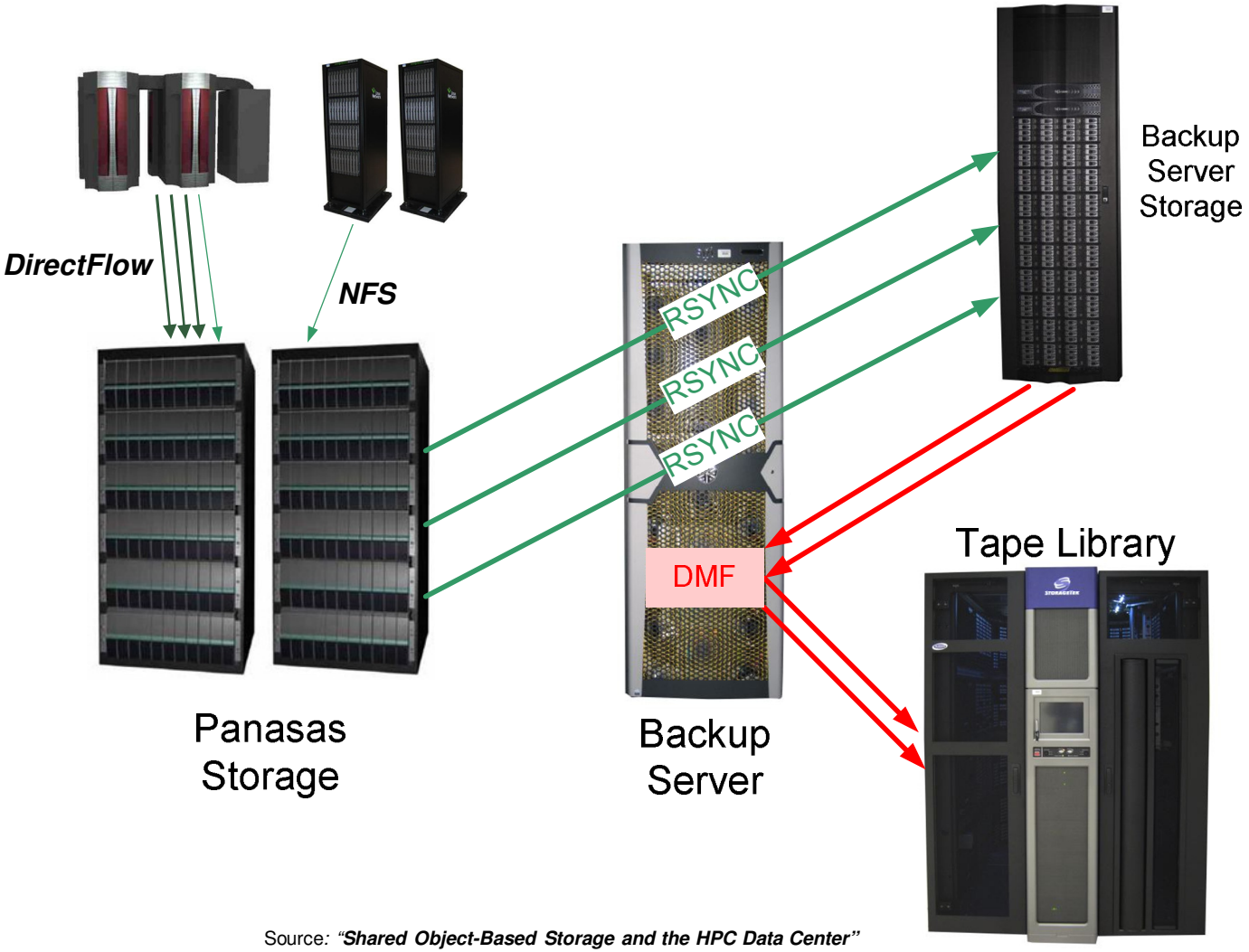
CAG & IDS, Locations in USA



- **Profile**
  - Use of HPC for design of commercial aircraft , space and communication and defense weapons systems
- **Challenge**
  - Deploy CAE simulation software for improvements in aerodynamic performance, reductions in noise, etc.
  - Provide HPC cluster environment to support 1000's of users for CFD (Overflow; CFD++; FLUENT), CSM (MSC.Nastran; Abaqus; LS-DYNA), and CEM (CARLOS)
- **HPC Solution**
  - 8 x Linux clusters (~3600 cores); 2 x Cray X1 (512 cores)
  - Panasas PanFS, 112 storage systems, > 900 TBs
- **Business Value**
  - CAE scalability allows rapid simulation turn-around, and enables Boeing to use HPC for reduction of expensive tests



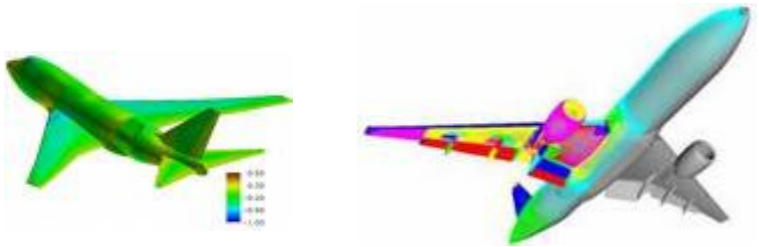
Panasas for Boeing HPC



Source: *"**Shared Object-Based Storage and the HPC Data Center**"*
Jim Glidewell/Claude Asher, High Performance Computing, Enterprise Storage and Servers, Boeing, Supercomputing Conference 2007

Panasas for Boeing HPC

- Panasas is meeting the high-performance storage requirements for Boeing's HPC facility:
  - Simple installation and easy Admin management
  - Superior DirectFlow performance and more than adequate NFS performance
  - Industry-leading post-sales support
  - Users are far more productive with quicker job turn-around
  - Shared common data storage has reduced data duplication and contained growth

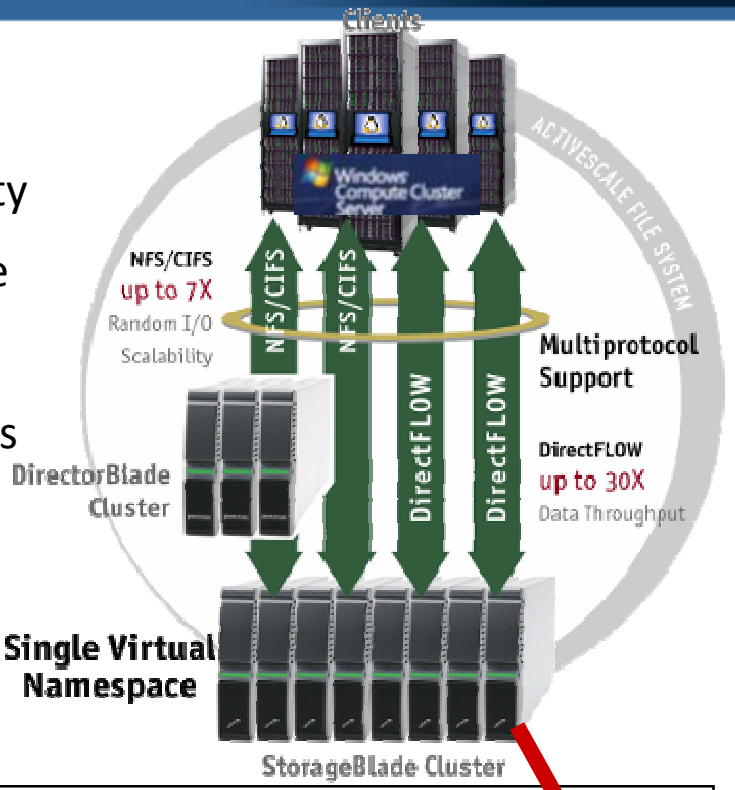


Source: *"**Shared Object-Based Storage and the HPC Data Center**"*
Jim Glidewell/Claude Asher, High Performance Computing, Enterprise Storage and Servers, Boeing, Supercomputing Conference 2007

Panasas Parallel File System and Storage

Parallel File System and Storage *Appliance*

- Parallel file system layered over an object-based architecture for scalability, reliability, manageability
- Panasas combines a PFS with architecture-aware storage hardware for ease in implementation
- High Performance: 600 MB/s per shelf and scales



Panasas parallel client S/W DirectFLOW

- Supports Linux (many variants, no kernel mods)
- Also multi-protocol support: NFS, CIFS, pNFS

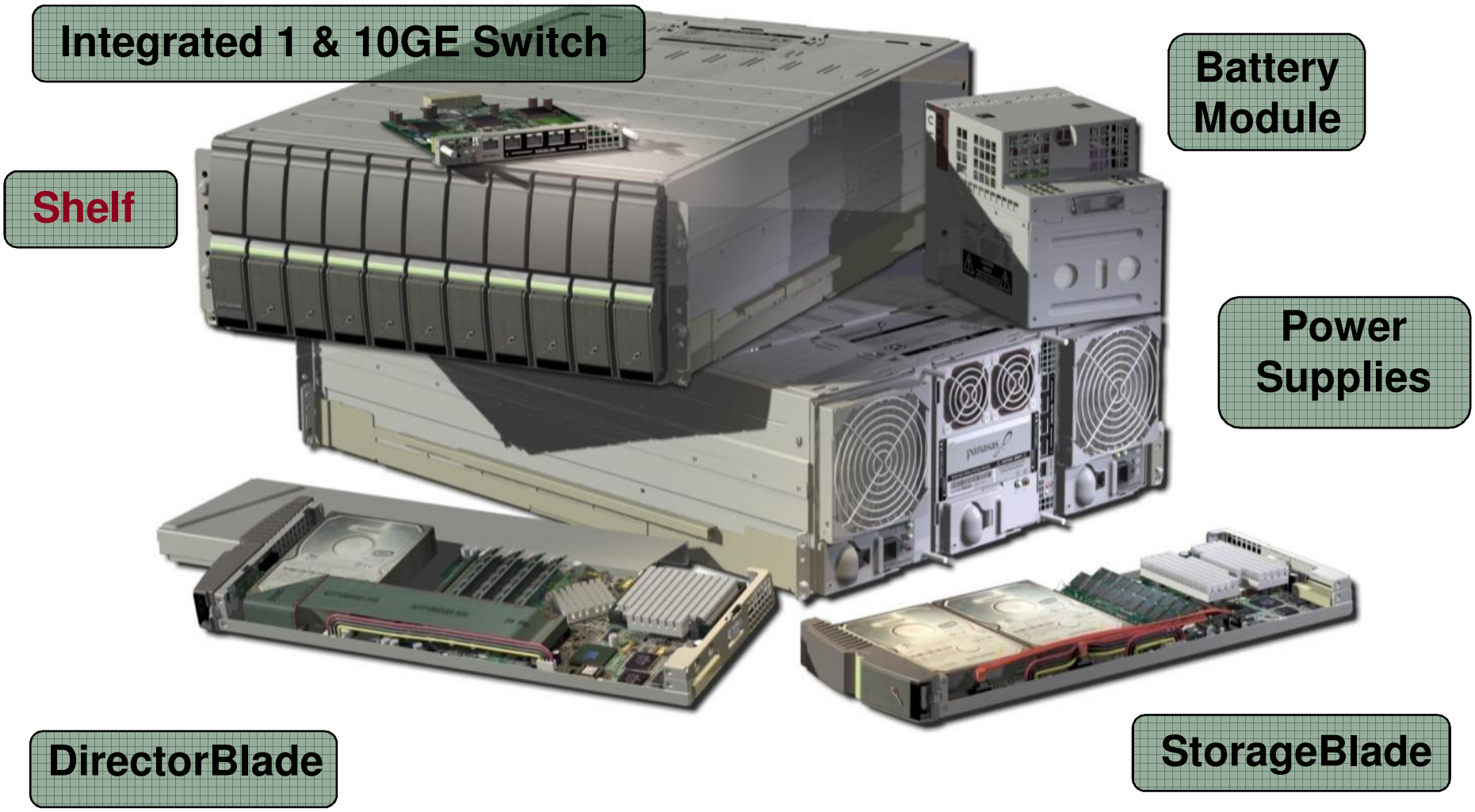
Panasas technology alliances

- ISVs with parallel CAE
- OEM Resellers: Bull, SGI, Dell
- Networking: Cisco; Force 10
- Intel ICR CERTIFIED
- Research organizations

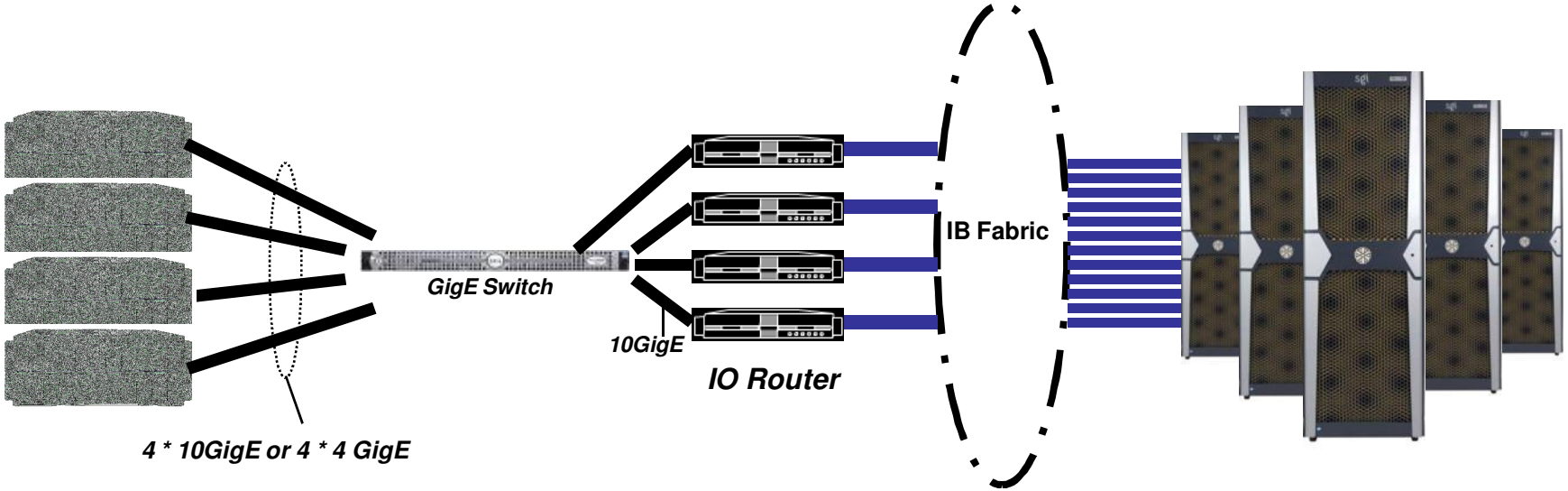
Description of the Blade-based Shelf

- Blade: disks, CPU, and memory
- 4U enclosure 11 blades per shelf
- Capacity 10, 15, or 20 TB per shelf
- Up to 20GB cache per shelf
- Up to 3 Director metadata blades
- 350 MB/s (GE) or 600 MB/s (10GE)
- Up to 10 shelves (200 TB) per rack

Panasas Shelf Built on Industry-Standards



Panasas I/O Router for IB Connectivity



- The Panasas I/O router has Infiniband and 10GigE
- Linux OS with OFED 1.3
- Transfer rate is up to 600 MB/second per router
- Four IO-routers can handle the load of four shelves using 10GigE
- Load is balanced over the IO-routers
- If one IO Router fails then the remaining ones take over the load

Panasas Industry Leadership in HPC

US DOE: Panasas selected for *Roadrunner,* ~2PB file system – top of Top 500

- LANL $133M system for weapons research: www.lanl.gov/roadrunner

SciDAC: Panasas CTO selected to lead Petascale Data Storage Institute

- CTO Gibson leads PDSI launched Sep 06, leveraging experience from PDSI members: LBNL/NERSC; LANL; ORNL; PNNL; Sandia NL; CMU; UCSC; UoMI

Aerospace: Airframes and engines, both commercial and defense

- Boeing HPC file system; 3 major engine mfg; top 3 U.S. defense contractors

Formula-1: HPC file system for Top 2 clusters – 3 teams in total

- Top clusters at an F-1 team with a UK HPC center and BMW Sauber

Intel: Certified Panasas storage for range of HPC applications – *Panasas Now ICR*

- Intel is a customer, uses Panasas storage in EDA and HPC benchmark center

SC08: Panasas won 5 of the annual HPC Wire Editor's and Reader's Choice Awards

- Awards for roadrunner (3) including "Top Supercomputing Achievemei
- "Top 5 vendors to watch in 2009" | "Reader's Best HPC Storage Product"

Validation: Panasas customers won 8 out of 12 HPC Wire industry awards for SC08:

Boeing | Renault F1 | Citadel | Ferrari F1 | Fugro | NIH | PGS | WETA

Questions

panasas

For more information,
call Panasas at:

1-888-PANASAS
(US & Canada)

00 (800) PANASAS2
(UK & France)

00 (800) 787-702
(Italy)

+001 (510) 608-7790
(All Other Countries)

Thank You