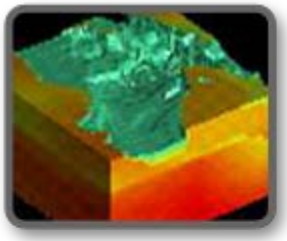
Tackling the Next Generation of HPC Challenges…

Delivering Scalability, Stability and Simplified Management for Massive Data, Storage and Compute

Peter Ungaro
President and CEO

Our Focus: Supercomputing

We build the world's fastest supercomputers to help solve "Grand Challenges" in science and engineering

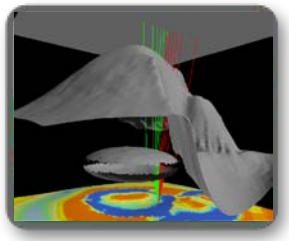


Earth Sciences
CLIMATE CHANGE
& WEATHER PREDICTION

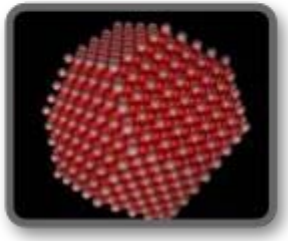
Life Sciences
PERSONALIZED MEDICINE &
IMPROVED BIOFUELS

Defense & National Security
WARFIGHTER SUPPORT,
THREAT PREDICTION &
STOCKPILE STEWARDSHIP

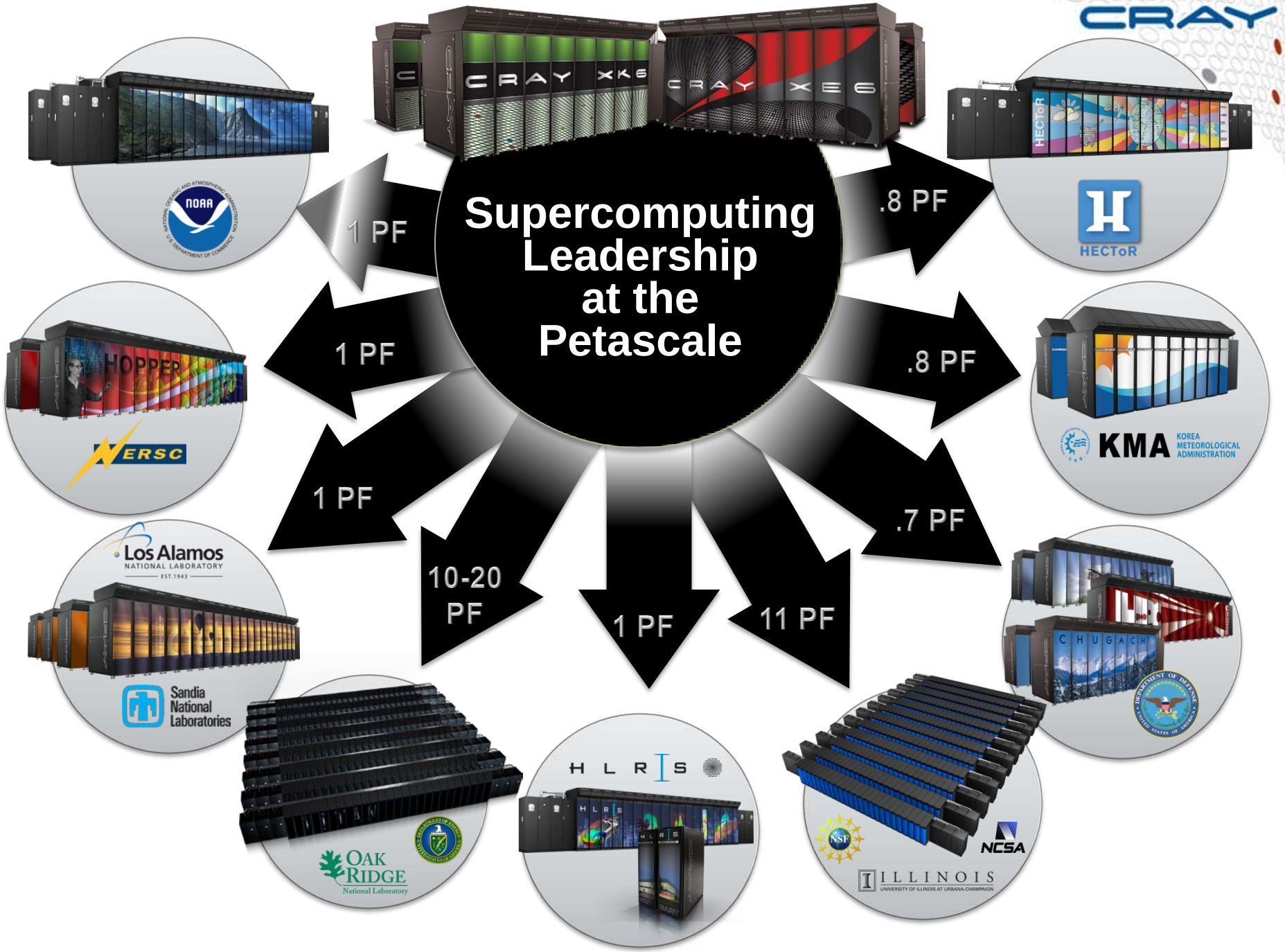
Computer-Aided Engineering
AIRCRAFT DESIGN,
CRASH SIMULATION &
FLUID DYNAMICS



Petroleum
SEISMIC IMAGING &
RESERVOIR SIMULATION



Scientific Research
NEW ENERGY SOURCES &
EFFICIENT COMBUSTION


CRAY
Supercomputing Leadership at the Petascale
CRAY XK6
CRAY XE6
NOAA
1 PF
HECToR
.8 PF
HOPPER
NERSC
1 PF
KMA
KOREA METEOROLOGICAL ADMINISTRATION
.8 PF
Los Alamos NATIONAL LABORATORY EST. 1943
1 PF
.7 PF
10-20 PF
1 PF
11 PF
Sandia National Laboratories
CHUGACH
TITAN
OAK RIDGE National Laboratory
HLRS
NSF
NCSA
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

The Faster the Simulation …the more Data Generated

Source: Eric Green, Director, National Institute of Health: NextGen 101 Workshop

Large Cray Storage Installations

Cray has deployed many of the World's highest bandwidth file systems

Site	Sustained IOR Bandwidth (GB/s)	Usable Capacity (PB)
Air Force Research Lab	266	1.7
Oak Ridge National Lab	240	>10
US Army (ERDC)	120	0.8
NERSC	70	>2
Cielo (LANL/Sandia)	140	2.5
University of Stuttgart (HLRS)	80	>2.5
NCSA Blue Waters*	1000	>25

* Deployment will be complete late 2012

It Takes a Village…

Architected by Cray

Components from the top storage partners

NetApp™

Mellanox TECHNOLOGIES

DataDirect NETWORKS

xyratex

SPECTRA SPECTRALOGIC.COM

LSI

Quantum®

panasas®

DELL™

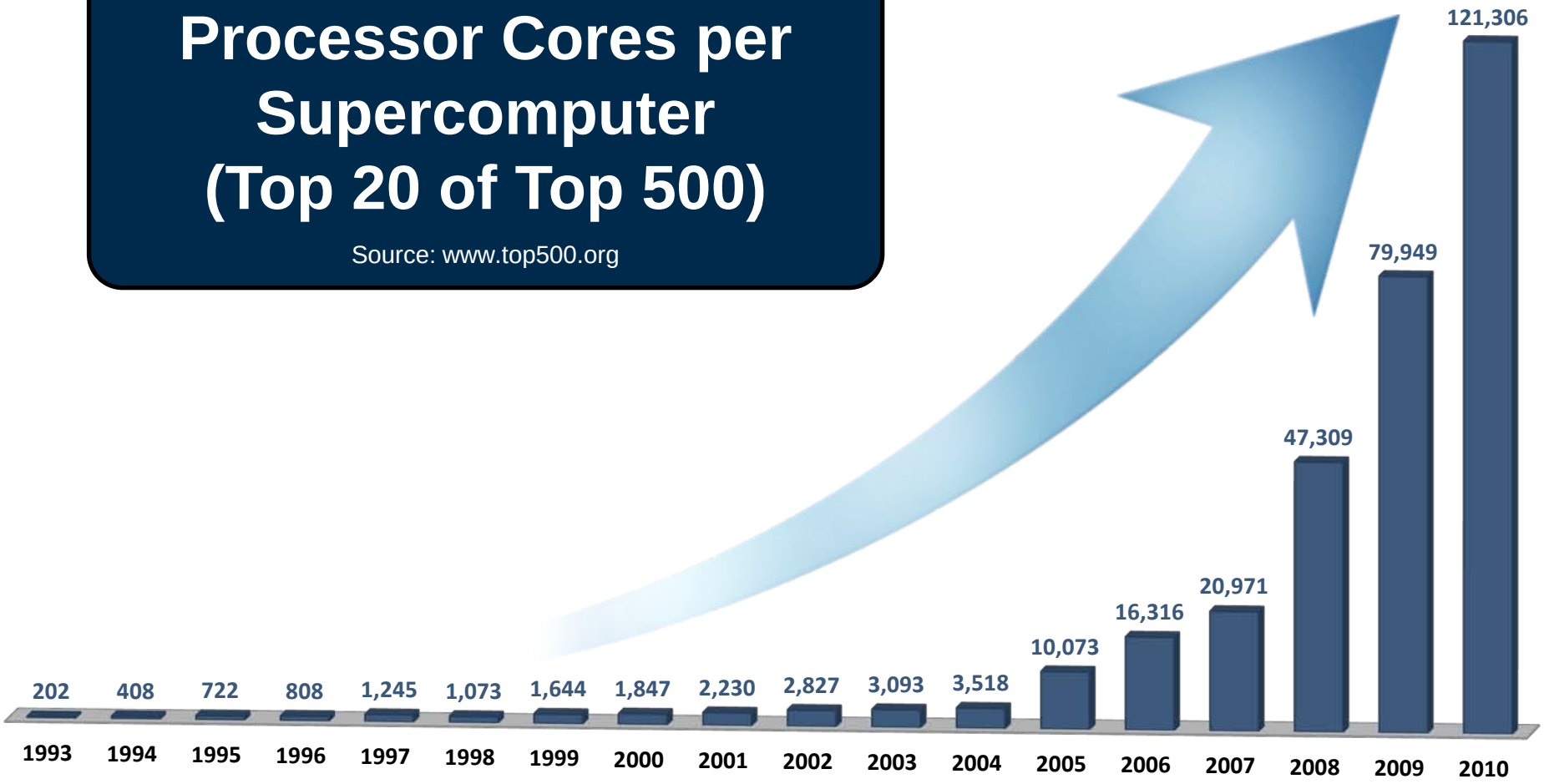
StorNext®

whamcloud

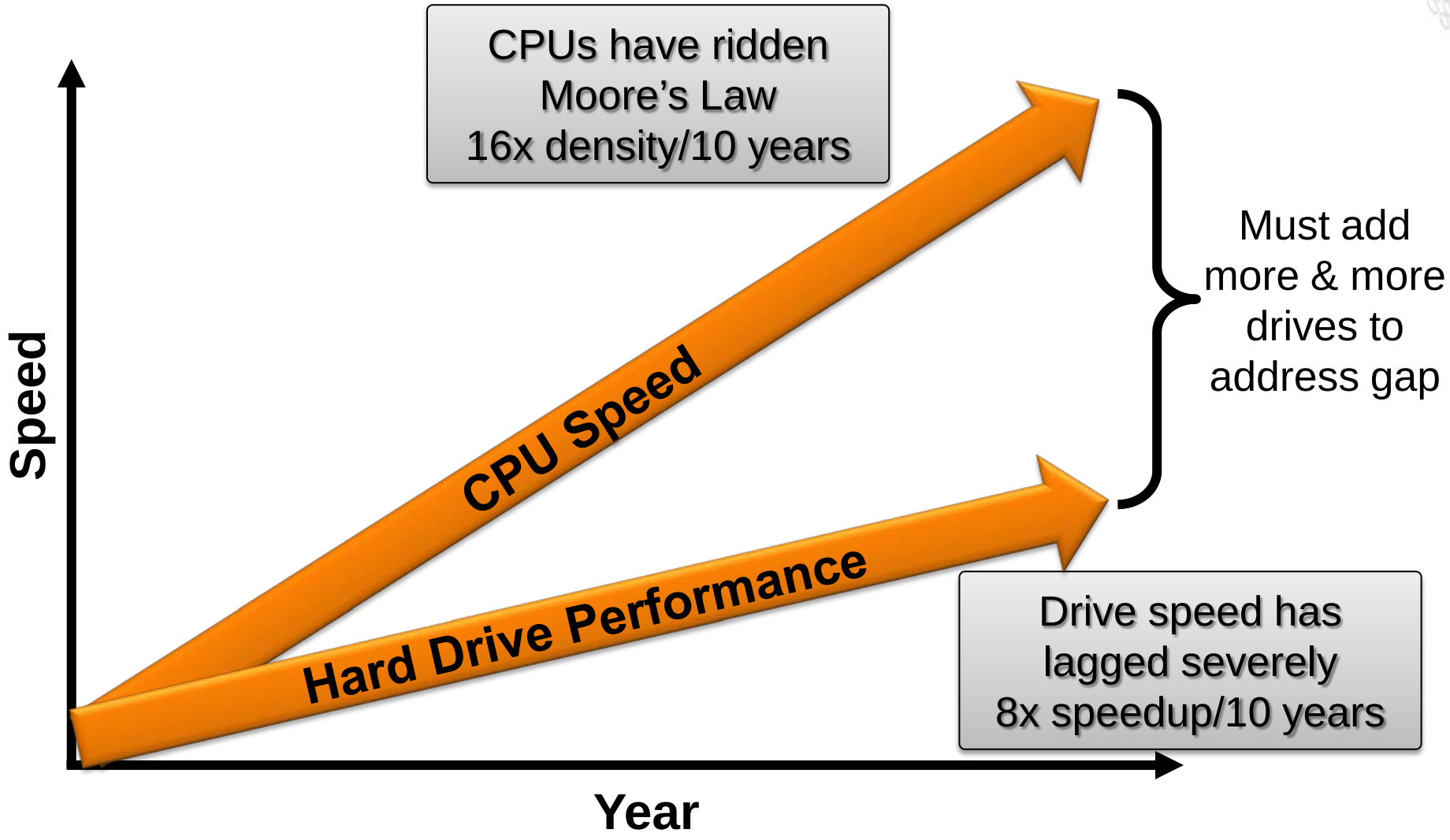
Scalability: Moore's Law[2]

Average Number of Processor Cores per Supercomputer (Top 20 of Top 500)

Source: www.top500.org



Moore's Law Doesn't Apply to Hard Drives!



Two Trends…Three Requirements

Compute	
	Annual Improvement
# Cores/ CPU	↗
Performance/ Core	→

Storage	
	Annual Improvement
Capacity	↗
Transfer rate	→
Seek time	→

Scalability

- *Scalable bandwidth & capacity expansion*
- *Modular flexibility and simplicity*

Reliability

- *Factory tested*
- *Fast failover, fast repair*

Visibility

- *Instant full system monitoring*
- *Detailed drill down*

Facing the Reality of 10,000 Spinning Disks

HPC storage systems must tolerate continuous failure

Design for Resilience	Manufacture for Reliability	Handle Errors Transparently
Designed for High Availability	Thorough burn-in from components to entire rack	Automatic failover of key sub-systems
No single point of failure	Rigorous testing at scale	Minimal impact and duration of failure

lustre®

OpenSFS®
Open Scalable File Systems, Inc.

EOFS
European Open File System

Insuring the Future of Lustre…

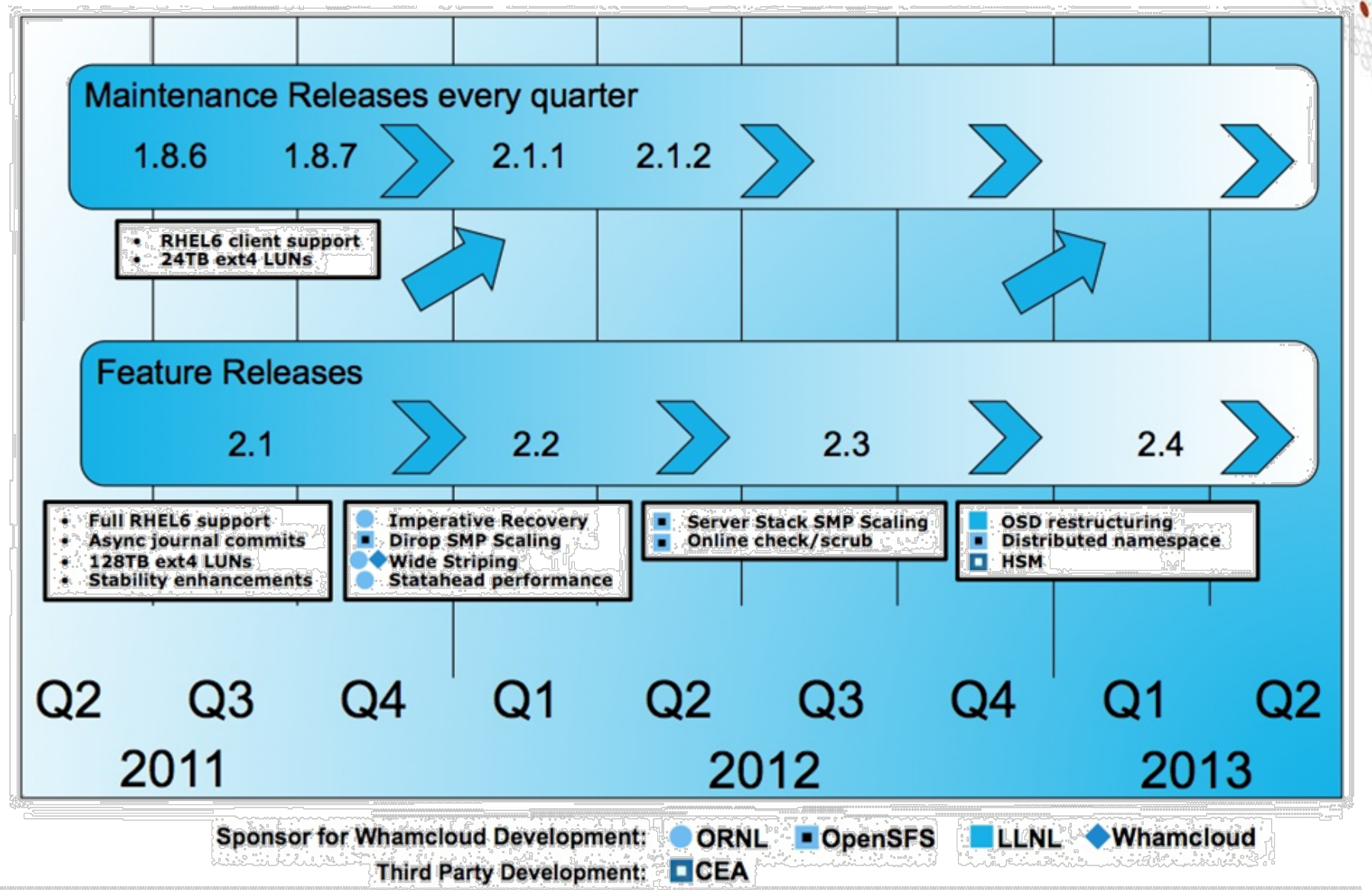
OpenSFS is a vendor neutral, member supported non-profit organization bringing together the open source file system community for the high performance computing sector

Our mission is to aggregate community resources and be the center of collaborative activities to ensure efficient coordination of technology advancement, development and education

The end goal is the continued evolution of robust open source file systems for, and under the control of, the HPC community

Promoters: *Cray*, DDN*, Oak Ridge*, LLNL*, Xyratex (* = co-founders)*
Adopters and Supporters: *Indiana University, NetApp, Sandia, Terascala, TACC, SGI, Whamcloud, RAID Incorporated*

Lustre Community Roadmap



Source:http://wiki.whamcloud.com/display/PUB/Community+Lustre+Roadmap

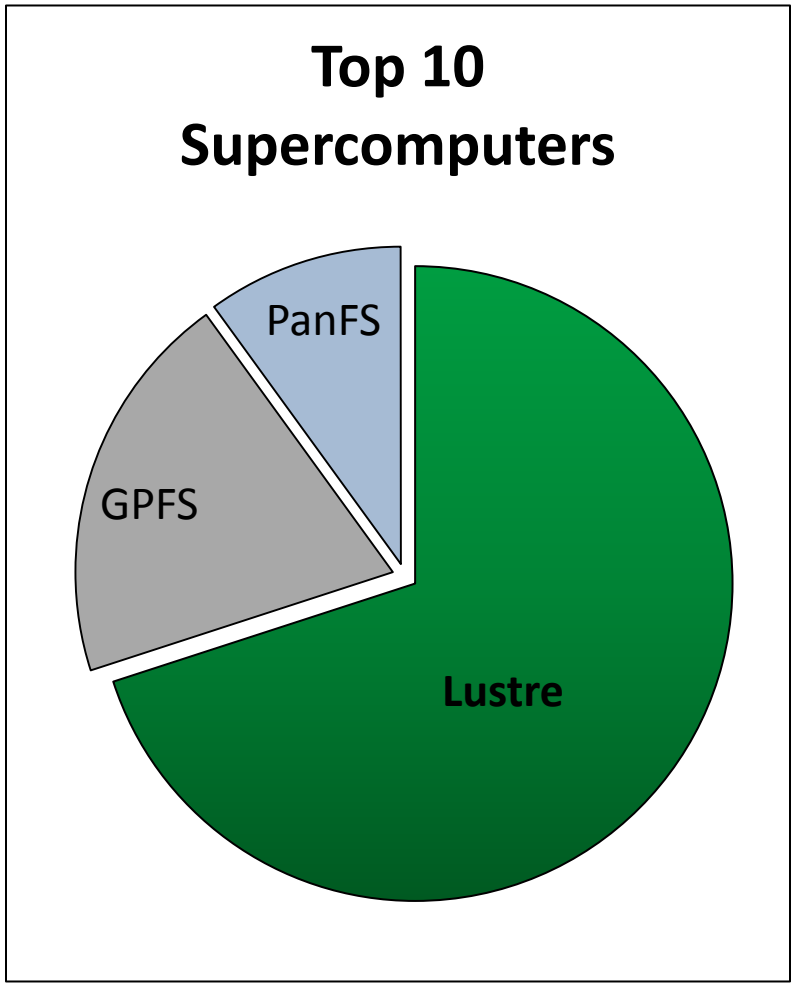
…Still Lots Of Work In Front Of Us

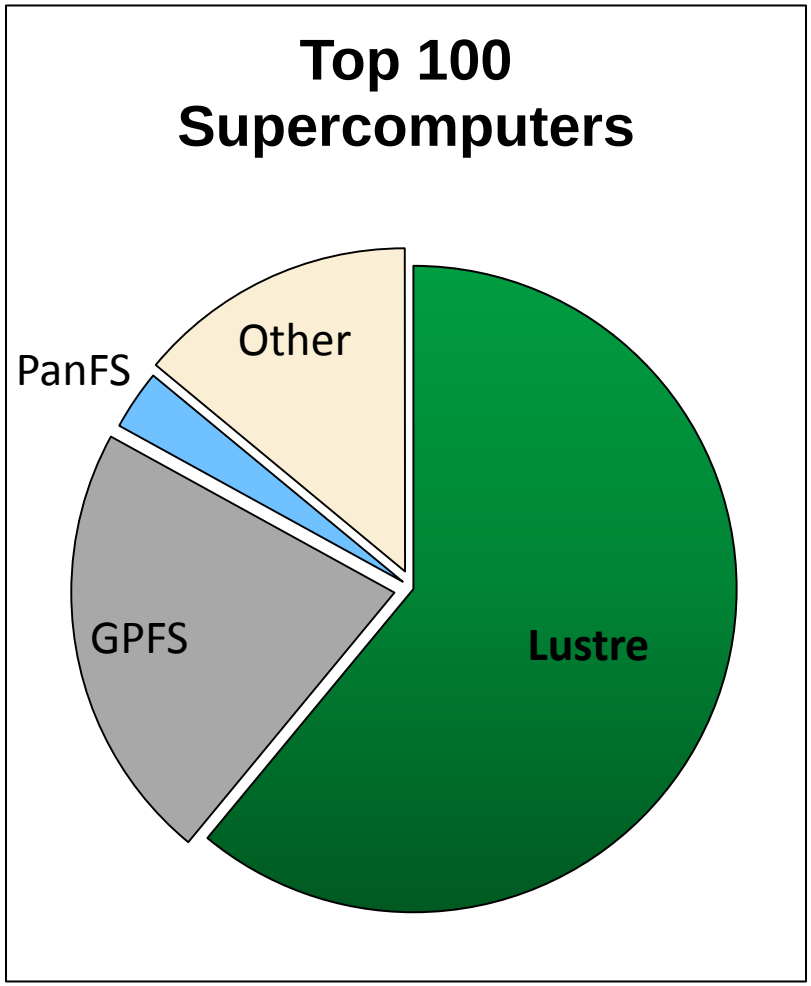
- **Metadata Scalability**
  - Can be bottleneck
  - Not optimized for small files
- **Quality of Service**
- **HSM**
- **End-to-End Data Integrity**
- **Quotas**
- **Wide Striping**

Funded & On-Going Projects	
Feature	**Impact**
Wide Striping	File Size
Imperative Recovery	Usability
Statahead	MDS
Parallel Directory Ops	MDS
SMP Affinity	MDS
Online file system check	Reliability
DNE Remote Directories	MDS
HSM	Usability
Object Storage Devices	Backing Store
DNE Striped Directories	MDS
OSD Quotas	Backing Store

Lustre Momentum

Proven scalability and performance at the worlds largest installations





Cray DVS supports all three file systems – full access to data anywhere

Holistic Approach to Petascale Computing…

We must deliver petascale, not just petaflops, solutions

(P.S. Ditto for Exascale)

Two Sides of the "Big Data Problem"

1

2

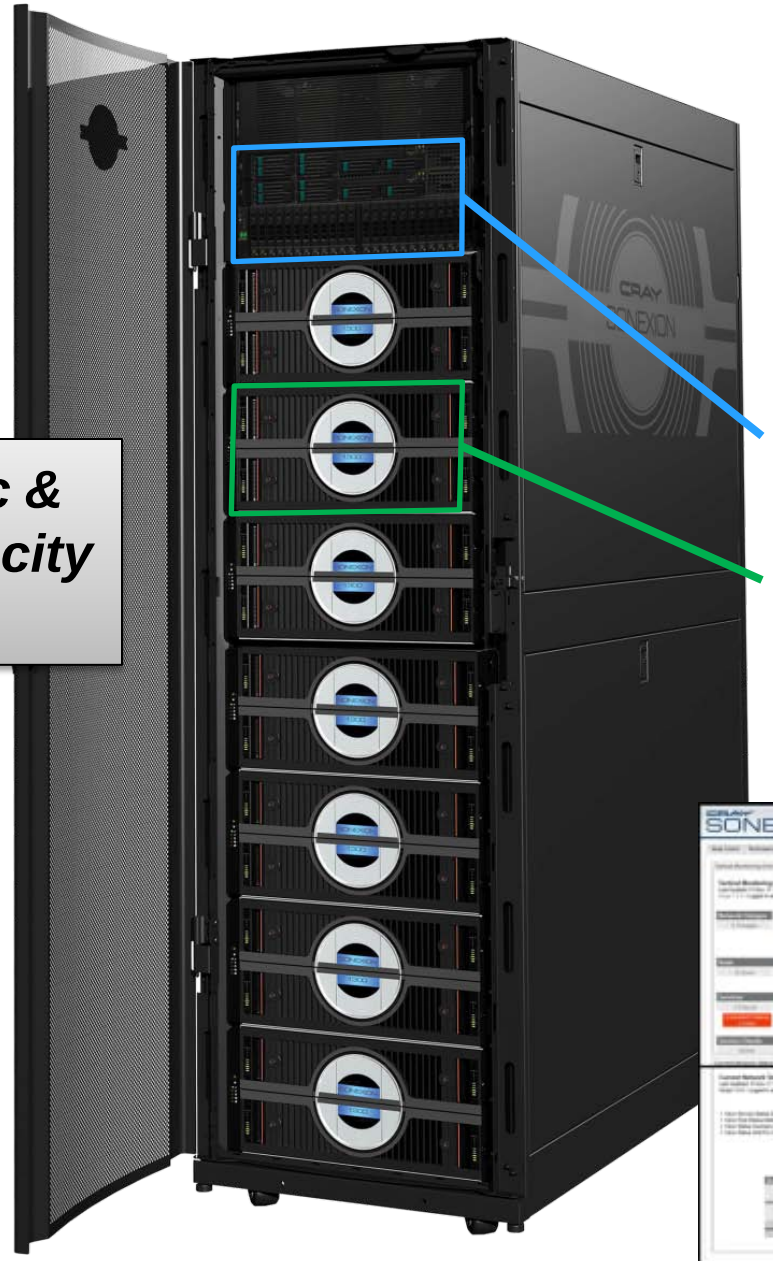
Huge Data Storage with High Performance I/O

- *Sonexion: Integrated Lustre storage solution*

Data Analytics

- *uRiKA: Graph Appliance for Relationship Analytics*

CRAY SONEXION: Scalability, Reliability & Visibility



> 20GB/sec & 1.2 PB Capacity per rack

3 Simple, Modular Components

Rack

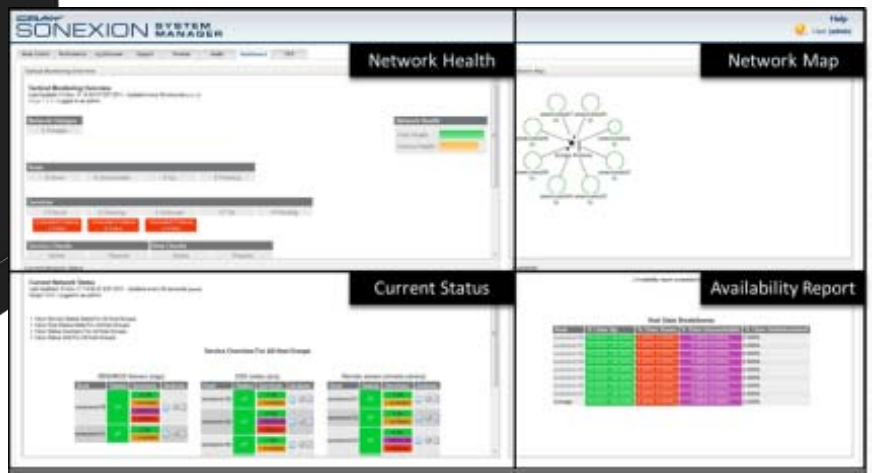
- Preconfigured power, cabling & switches

Metadata Management Unit

- Supports up to 6 billion files

Scalable Storage Unit

- 3 GB/sec sustained
- Performance scales with capacity
- Lustre can scale up to 64 PB



uRiKA: Big Data Graph Appliance for Relationship Analytics

Discover Unknown and Hidden Relationships in Big Data

- Relationship Warehouse supporting Inferencing/Deduction, Pattern-based queries and Intuitive Visualization

Perform Real-time Analytics on Big Data Graph Problems

- High-performance, Graph Appliance with large shared-memory, massive multi-threading and scalable I/O

Realize Rapid Time to Value on Big Data Solutions

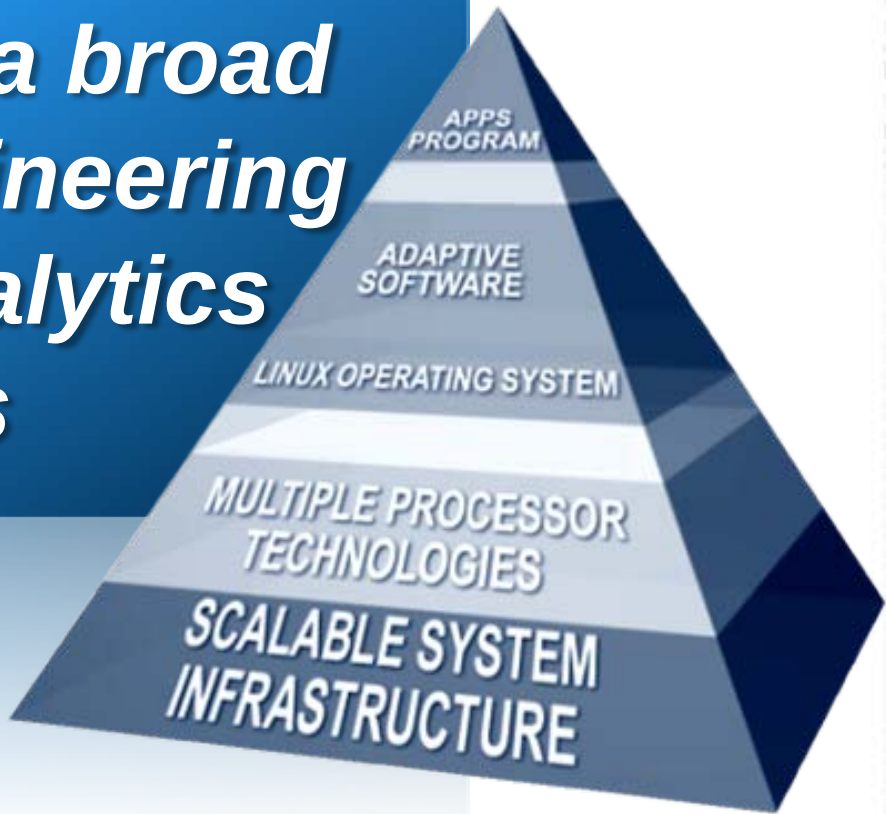
- Ease of Enterprise adoption with industry-standards, open-source software stack enabling reuse of existing skillsets and no lock-in

Pulling it all together….
Blue Waters: a Hybrid, Balanced Petascale System

Cray System & Storage cabinets:	• >300
Compute nodes:	• >25,000
System Memory:	• >1.5 Petabytes
Usable Storage; Bandwidth:	• >25 Petabytes; >1TB/sec
Peak performance:	• >11.5 Petaflops
Number of AMD processors; cores:	• >47,000; >380,000
Number of NVIDIA GPUs; cores:	• >3,000; > 1.5M

Our Vision: Exascale

Build a world-class supercomputer that enables transformational computing across a broad set of science, engineering and advanced analytics applications



Exascale Metrics and Technologies

	Today	2015	2018	2020
Max Nodes (M)	0.1	0.1	1	1
Max Concurrency (M)	1	10	100	1000
Memory (DRAM) (PB)	4	10	30	60
MTTA Interrupt (days)	1	1	0.5	0.25
Dump Memory (seconds)	2000	1000	600	300
Scratch File system (PB)	100	300	900	3000
Peak IO Burst (TB)	2	10	50	100
Metadata Transactions (M/per second)	0.1	1	10	100
Storage Application Visible interrupt (days)	20	18	16	14

Spinning Disks Resiliency Rebuilds

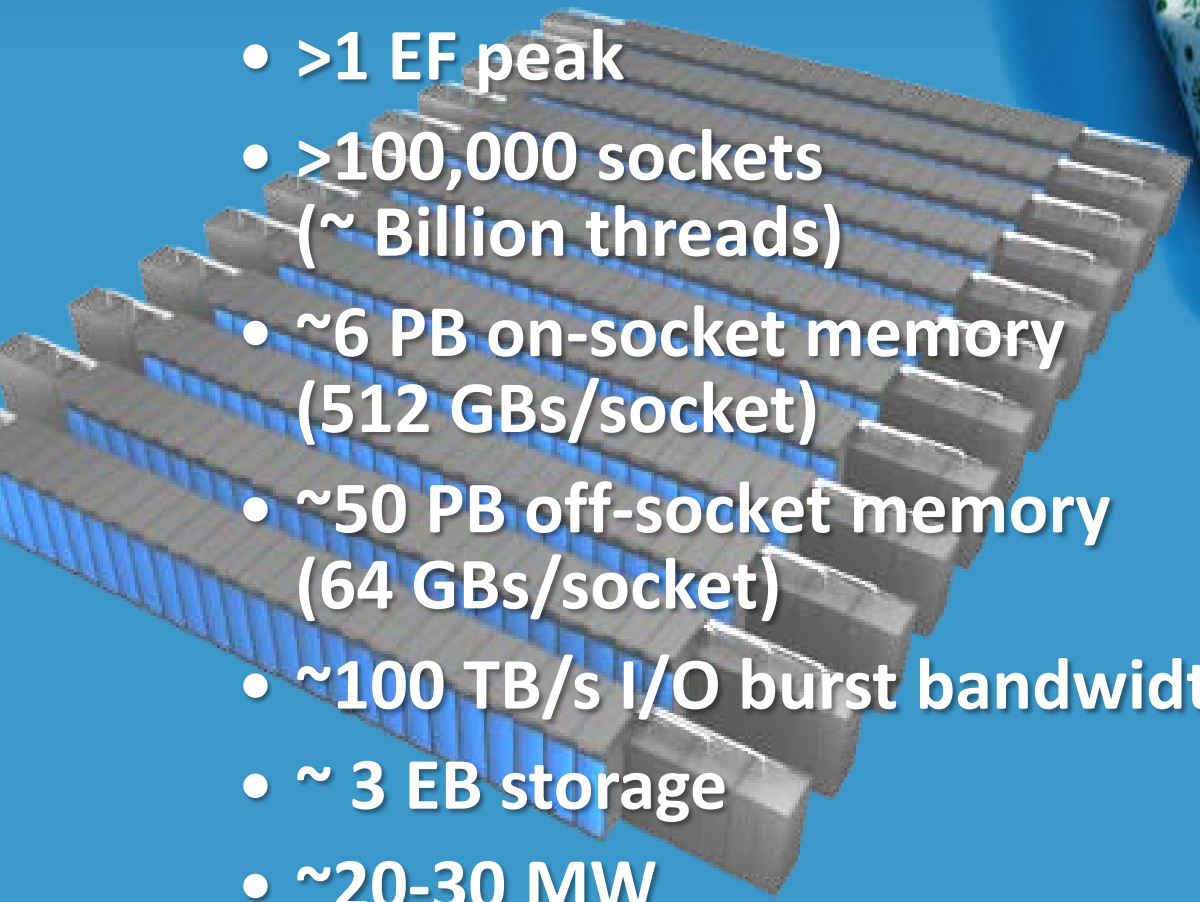
SSD or other high rate technology

Parallel Metadata + Innovative Technologies

Application Awareness and Resiliency Middleware SW

Concept Exascale System

- ~ 250 cabinets
- ~12-14 TF processor
- ~5 PF per cabinet
- >1 EF peak
- >100,000 sockets
  (~ Billion threads)
- ~6 PB on-socket memory
  (512 GBs/socket)
- ~50 PB off-socket memory
  (64 GBs/socket)
- ~100 TB/s I/O burst bandwidth
- ~ 3 EB storage
- ~20-30 MW



Enabling Simulation and Analytics

Computation

(Programmability)

Data Storage

(Enablement)

Data Analytics

(Relationships)

Adaptive Supercomputing

The HPC Evolution…

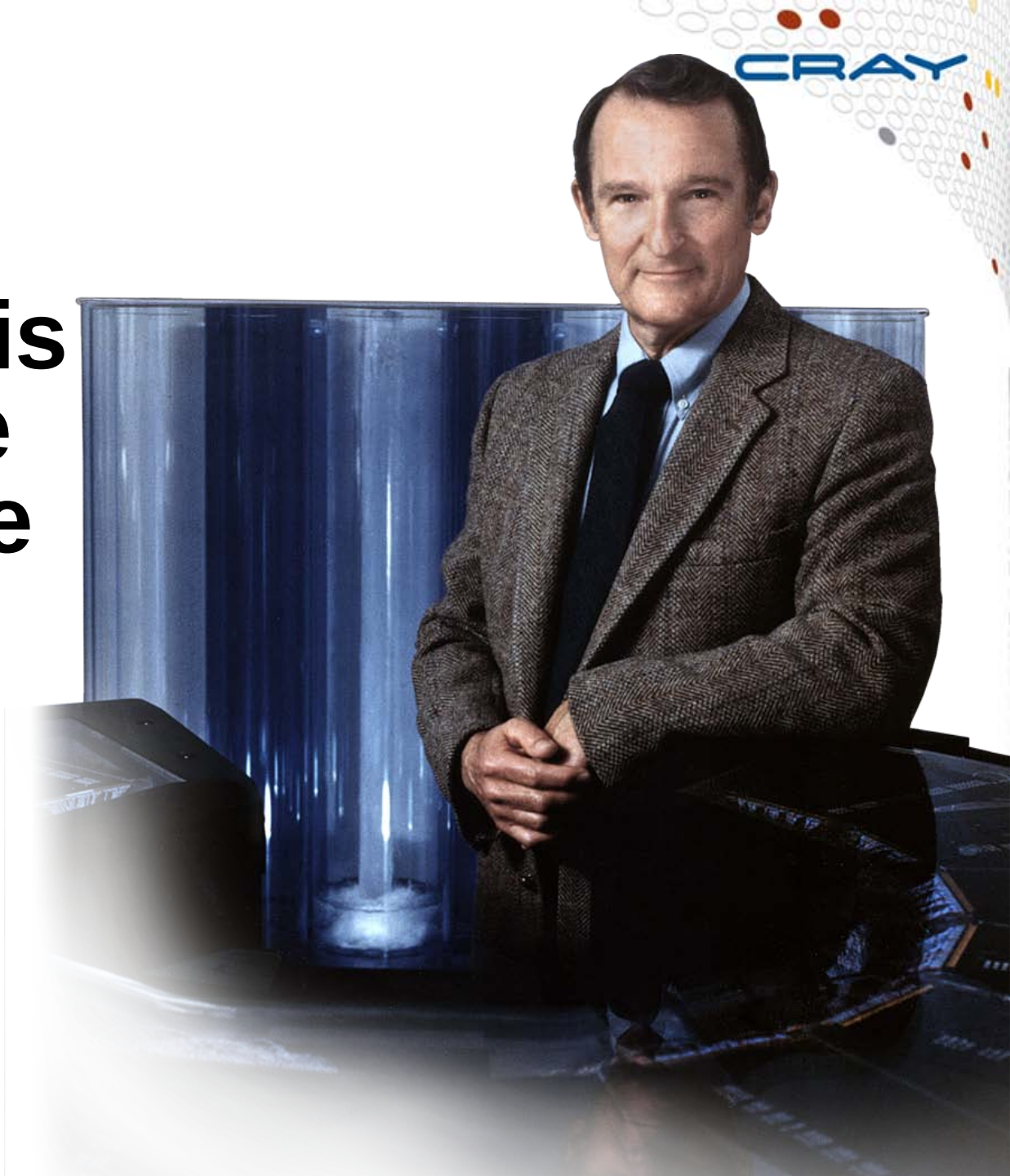
Existing architectures are being stressed as technology trends and scalability needs are accelerating

Scalable and portable open source file systems are the best path to engage the community & deliver on future requirements

A tightly-integrated, holistic approach to the HPC environment is required…
across massive compute & massive storage

"The future is seldom the same as the past"

Seymour Cray
June 4, 1995



Thank You!

CRAY
THE SUPERCOMPUTER COMPANY