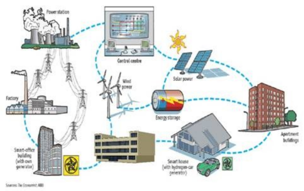
HPC and BigData solutions in Research

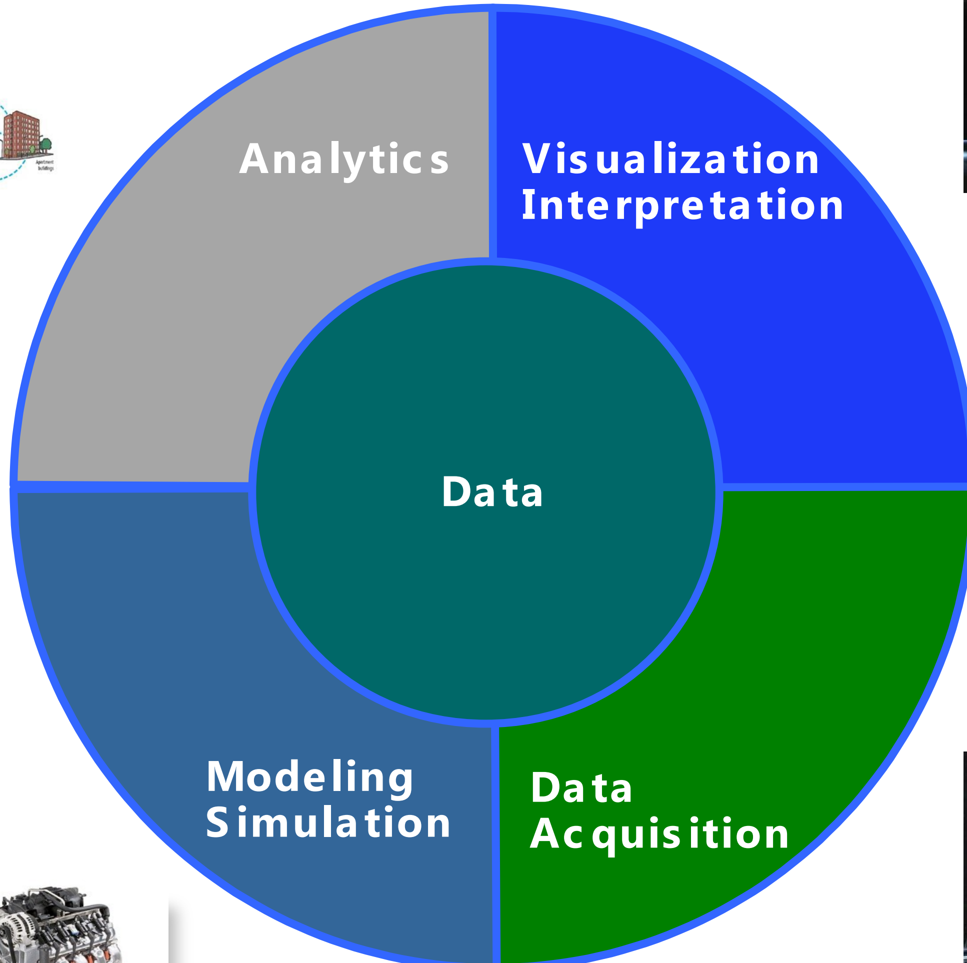
Marco Briscolini
marco_briscolini@it.ibm.com

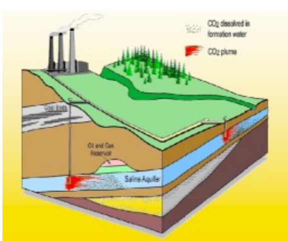
Workshop ICT – INAF 2014 - Pula

© 2013 IBM Corporation

End to End Analysis



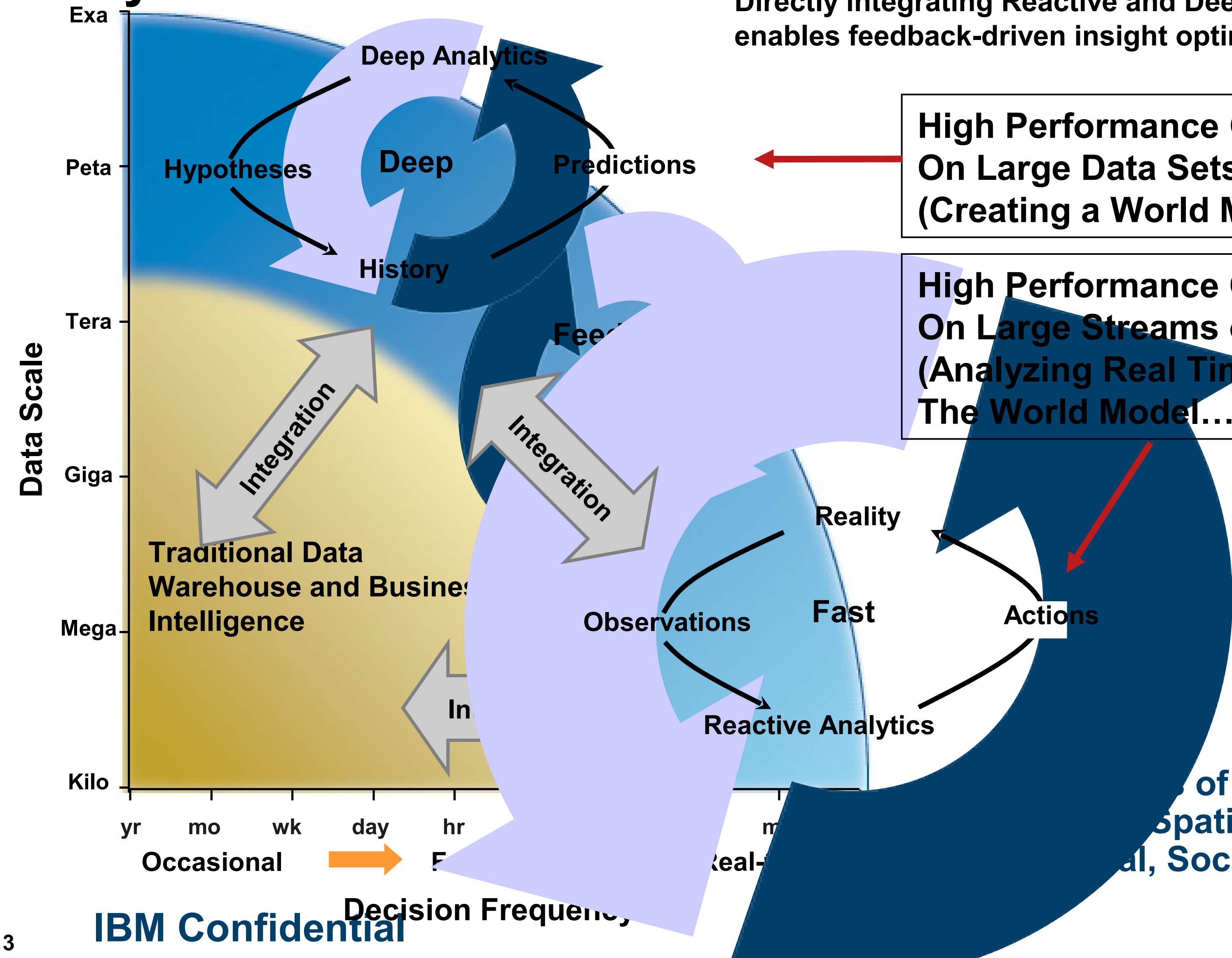




Massive Scale Data and Compute

Maximum Insight Requires Combining Deep and Reactive Analytics

Directly integrating Reactive and Deep Analytics enables feedback-driven insight optimization



High Performance Computing On Large Data Sets (Creating a World Model …Context)

High Performance Computing On Large Streams of Data (Analyzing Real Time against The World Model… Context)

of Streams of Data:
Spatial, Video/Image,
al, Social Network, etc

Stream Computing

- Stream computing for real time analytics on data in motion
  - It’s not just about the need for immediate action
  - Stream computing can be much more effective & efficient

Data in Motion

Data at Rest

Real Time Analytics
Drinking from the fire hose

- Huge volumes of data from a variety of sources
- Process & filter data without storing it
- Only store data which is of value
- Only alert users to data which is of interest
- Act on Information as it’s happening

Future Data-Centric Systems Challenges

	Traditional Systems	Data-Centric Systems
Computation Requirements	▪ Pre Planned Computation balance ▪ Static numerical computations & simulations	▪ Dynamic compute sensitive to data locale ▪ Data-driven analytics & graph computations
Communication Requirements	Mostly to neighbors	Each smart communication agent (processing/storage/memory)
Execution Environment	Static load balancing, algorithm driven	Data-Centric: Scheduling, data management and data curation
Computational Agents	Compute in cores	▪ Compute in cores ▪ Compute in memory ▪ Compute in network ▪ Compute in storage
Parallelism	Chip-based: threads, vectors, cores & nodes	Ubiquitous parallelism: computation grows to other systems components
Network Drivers	▪ 2D/3D stencil ▪ Coarse grained, point to point communication	▪ All-to-all ▪ Fine-grained communication ▪ Rich set of memory atomics
Network Topology	High diameter (Mesh/Torus)	Low diameter (Dragonfly, Fat tree)

IBM Data-Centric Design Principles

The onslaught of Big Data requires a composable architecture for big data, complex analytics, modeling and simulation. The DCS architecture will appeal to segments experiencing an explosion of data and the associated computational demands

Principle 1: Minimize data motion

- Data motion is expensive
- Hardware and software to support & enable compute in data
- Allow workloads to run where they run best

Principle 2: Enable compute in all levels of the systems hierarchy

- Introduce "active" system elements, including network, memory, storage, etc.
- HW & SW innovations to support / enable compute in data

Principle 3: Modularity

- Balanced, composable architecture for Big Data analytics, modeling and simulation
- Modular and upgradeable design, scalable from sub rack to 100's of racks

Principle 4: Application-driven design

- Use real workloads/workflows to drive design points
- Co-design for customer value

Architectural Elements

- Compute dense nodes
- Integrated compute and storage nodes for working set data and for data intensive work
- Common active network: IB/Ethernet, initially, moving to new IBM network

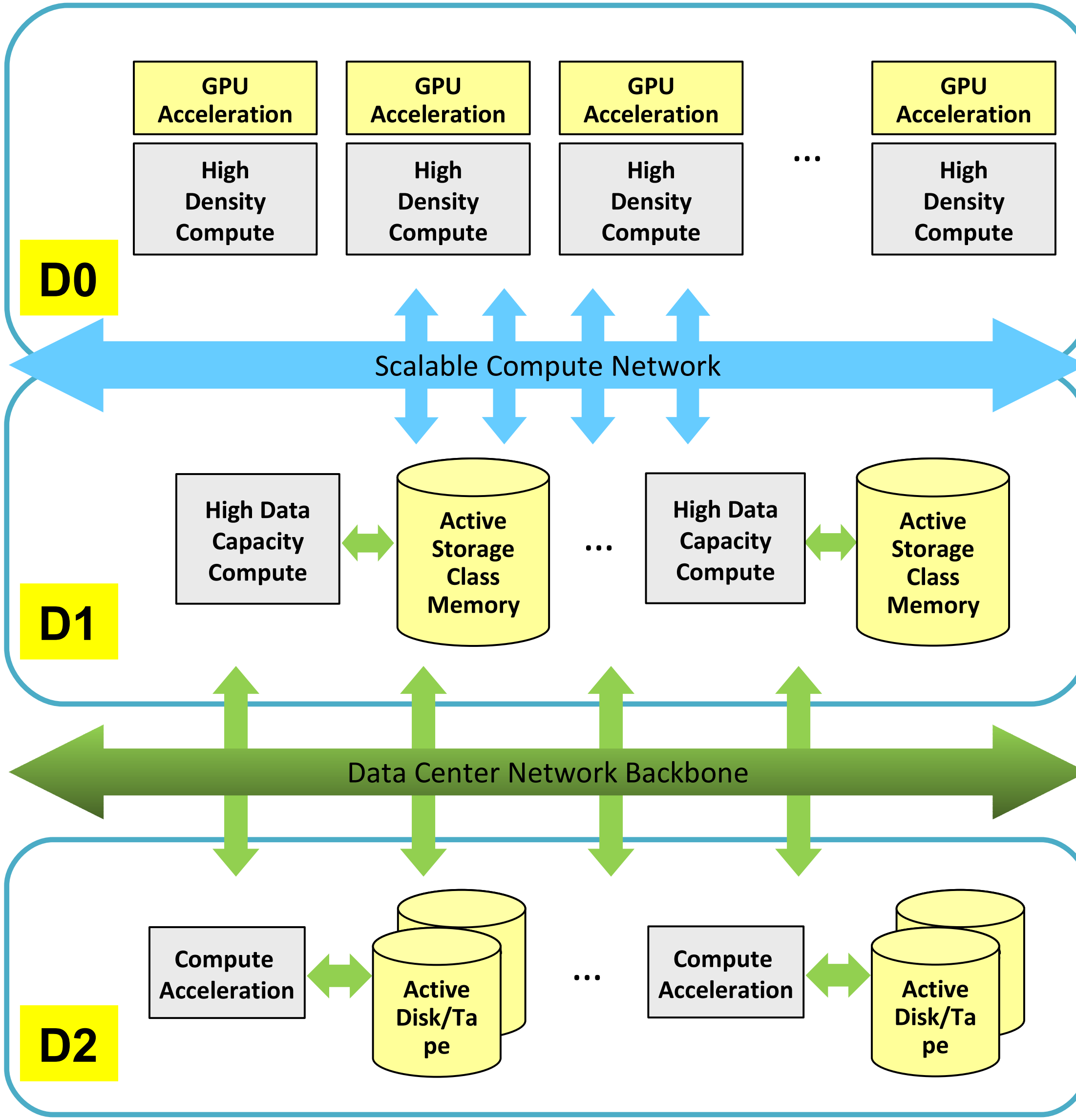
Data-Centric Deep Computing Architecture

IBM

Heterogeneous System
- Three Tier Optimization
  - Compute
  - Persistent Store
  - Data Reference
- Industry Std Data Center Connectivity
- Streaming capable

Modular Growth
- Introductory systems to 100's racks
- Non Disruptive Upgrades
- Capacity on Demand
- 'Cloud' like attributes
- Multi-workflow Capable

Data-Centric Model
- Optimized for Workflows
- Multiple Programming Models
  - Messages Passing
  - Threading
  - Global Address
  - Analytics Models
- Simulation and Modeling
- Analytics & Visualization
- Streaming

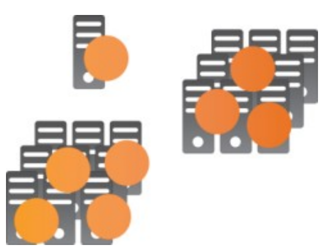


This HPC evolution is ushering in a corresponding evolution in HPC delivery.

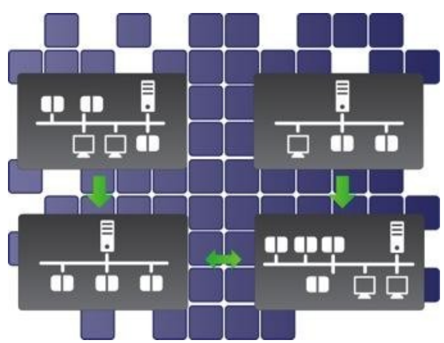
HPC Delivery Evolution

Scope of sharing

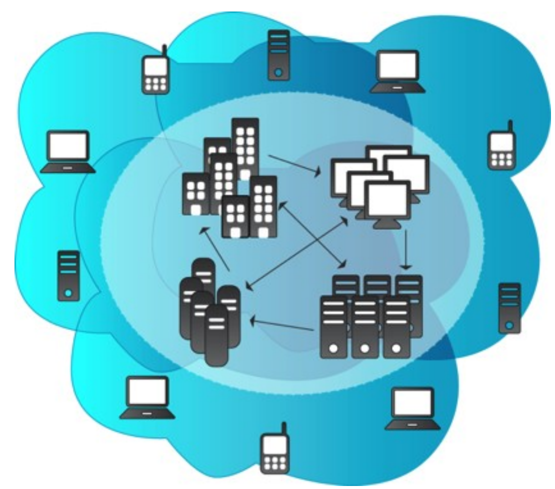
HPC Cluster
- Commodity hardware
- Compute / data intensive apps
- Single application/ user group



Grid
- Multiple applications or Group sharing resources
- Dynamic workload using static resources
- Policy-based scheduling

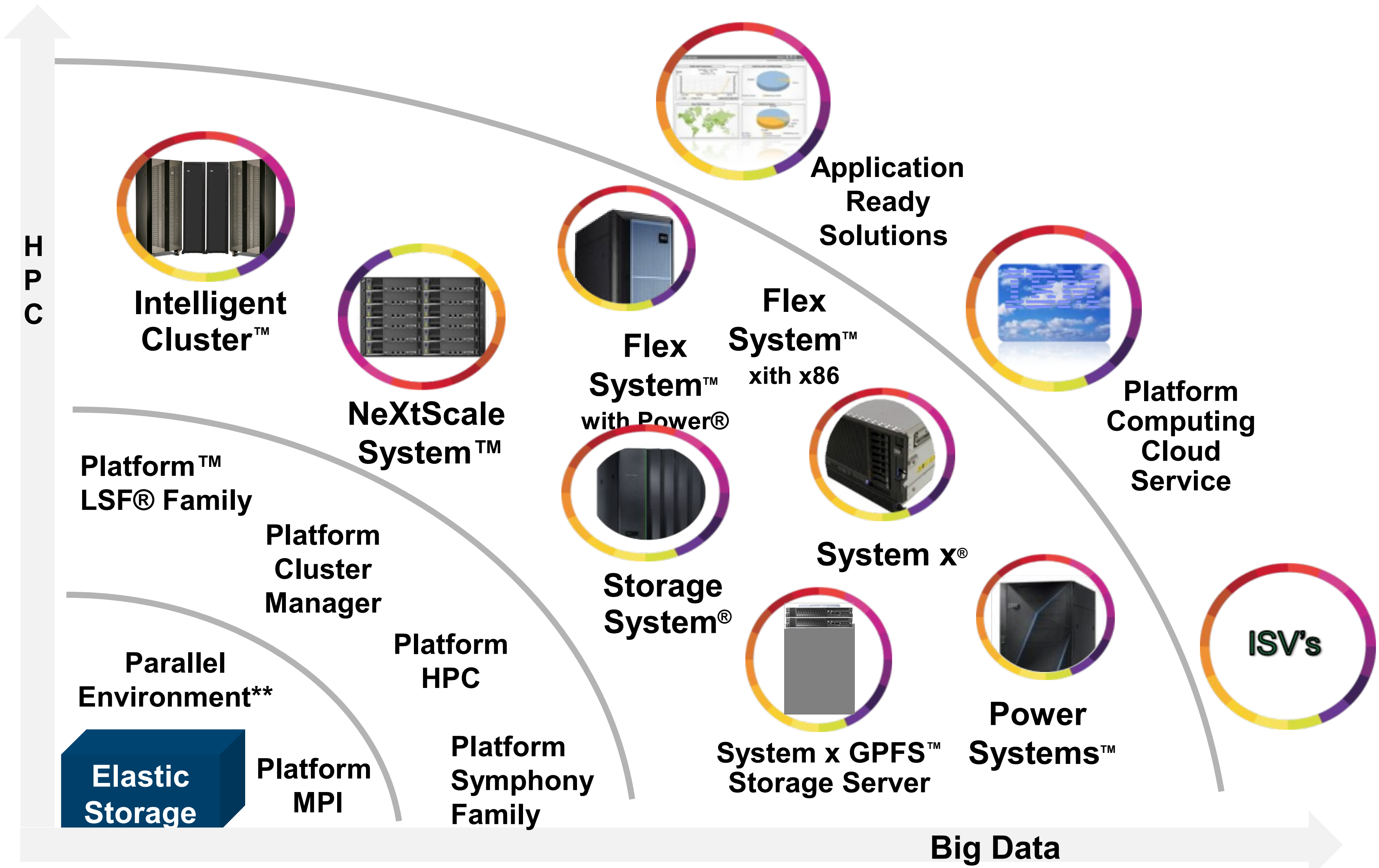


HPC Cloud
- HPC applications
- Enhanced self-service
- Dynamic HPC infrastructure: reconfigure, add, flex



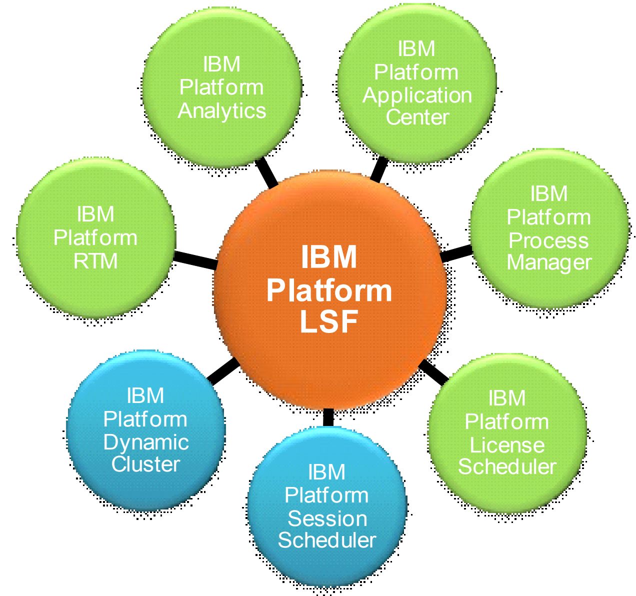
1992 2002 2012

Part of broad IBM Technical Computing portfolio



* LL maintenance only ** PE Power

IBM Platform LSF Product Family



Overview

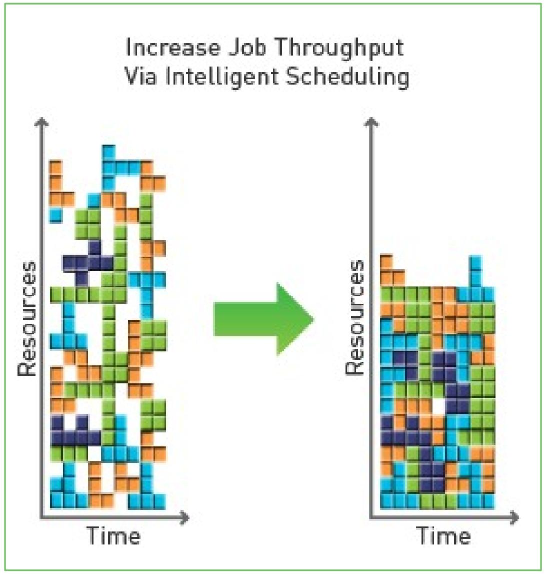
Powerful workload management for demanding, distributed and mission-critical high performance computing environments.

Key Capabilities

- **Complete**
  - **Advanced workload scheduling**
  - **Robust set of add-on features**
  - **Integrated application support**
- **Flexible**
  - **Heterogeneous platform support**
  - **Policy-driven automation**
  - **CLI, web services, APIs**
- **Powerful**
  - **Policy, resource and energy-aware scheduling**
  - **Resource consolidation for optimal performance**
  - **Advanced self-management**
- **Scalable**
  - **Thousands of concurrent users and jobs**
  - **Virtualized pool of shared resources**
  - **Flexible control, multiple policies**

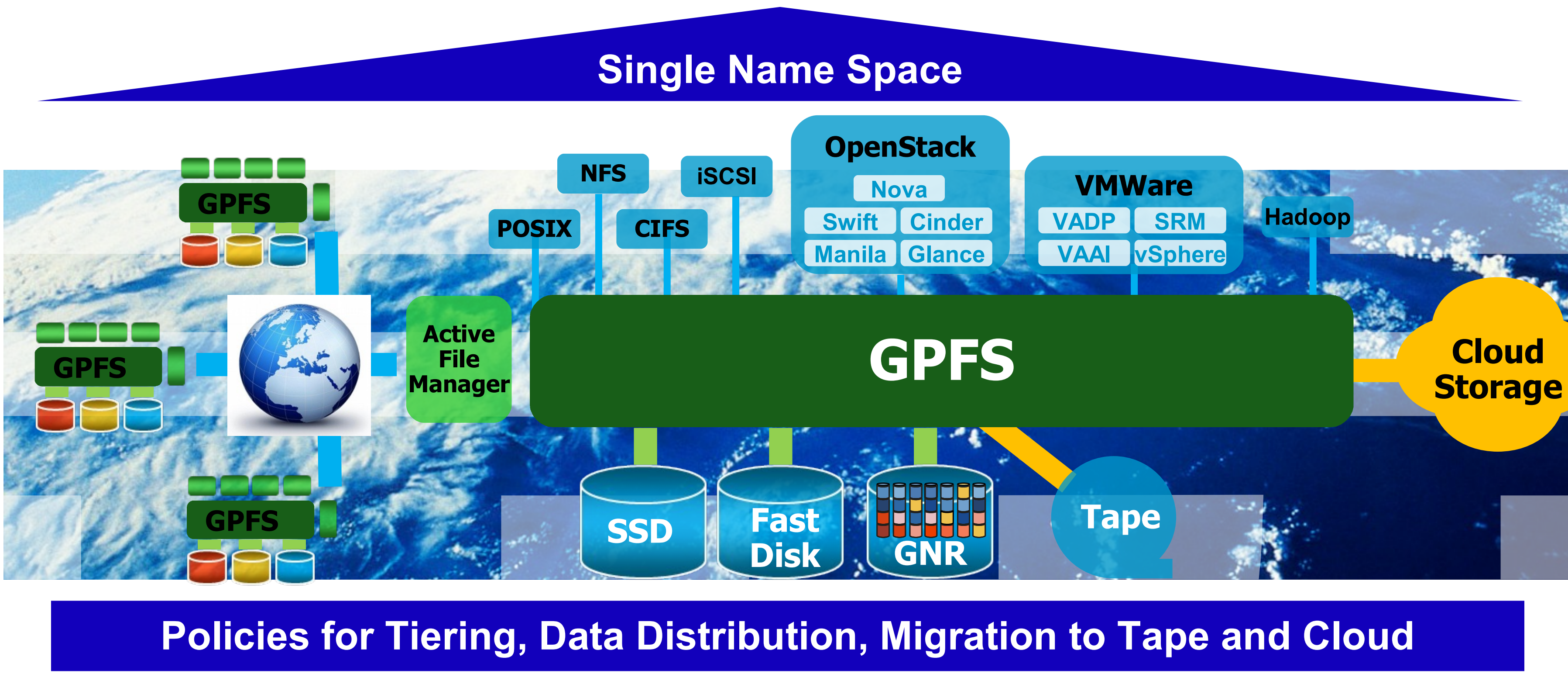
Client Benefits

- Optimal utilization: reduced infrastructure cost
- Robust capabilities: improved productivity
- High throughput: faster time to results

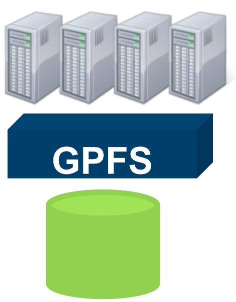


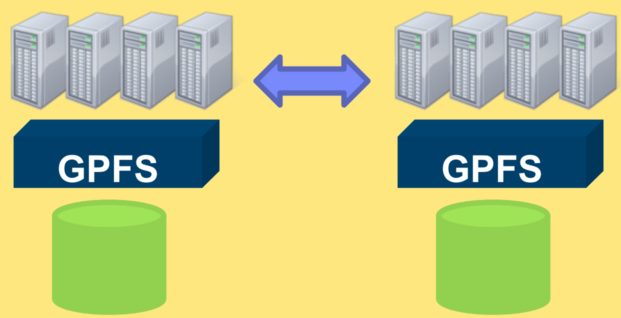
GPFS Cloud Data Plane Vision

- GPFS is a single scale-out data plane for the entire data center
- Unifies VM images, analytics, block devices, objects, and files
- Single name space no matter where data resides
- De-clustered parity - GPFS Native RAID (GNR)
- Data in best location, on the best tier (performance & cost), at the right time
- **All in software**



Evolution of the global namespace: GPFS Active File Management (afm)





GPFS introduced concurrent file system access from multiple nodes.

Multi-cluster expands the global namespace by connecting multiple sites

AFM takes global namespace truly global by automatically managing asynchronous replication of data

1993 **2005** **201**

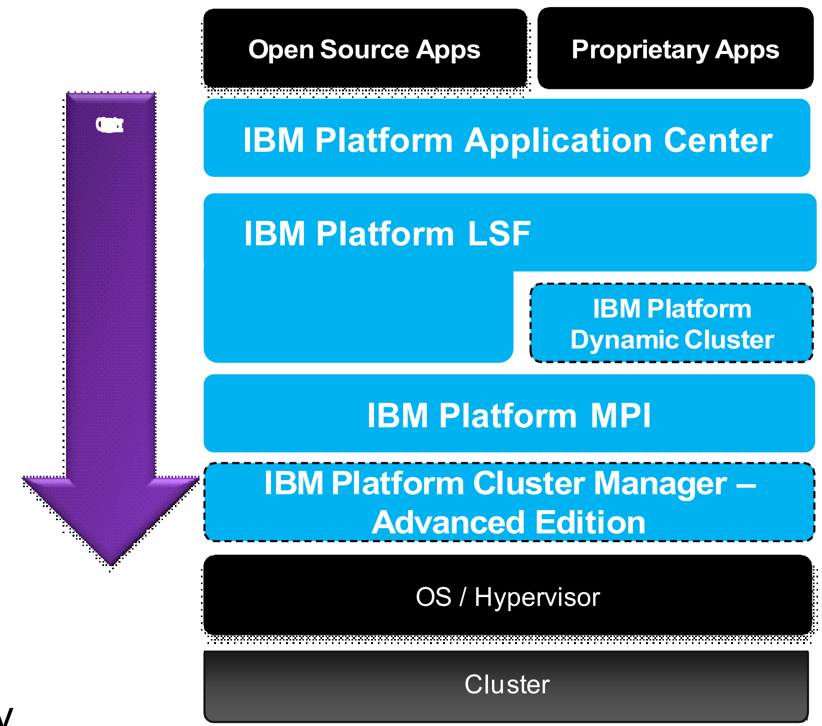
IBM Platform Computing Private Cloud Solutions

Overview

Innovative solutions for dynamic, flexible technical computing, big data & analytics cloud environments

Key capabilities

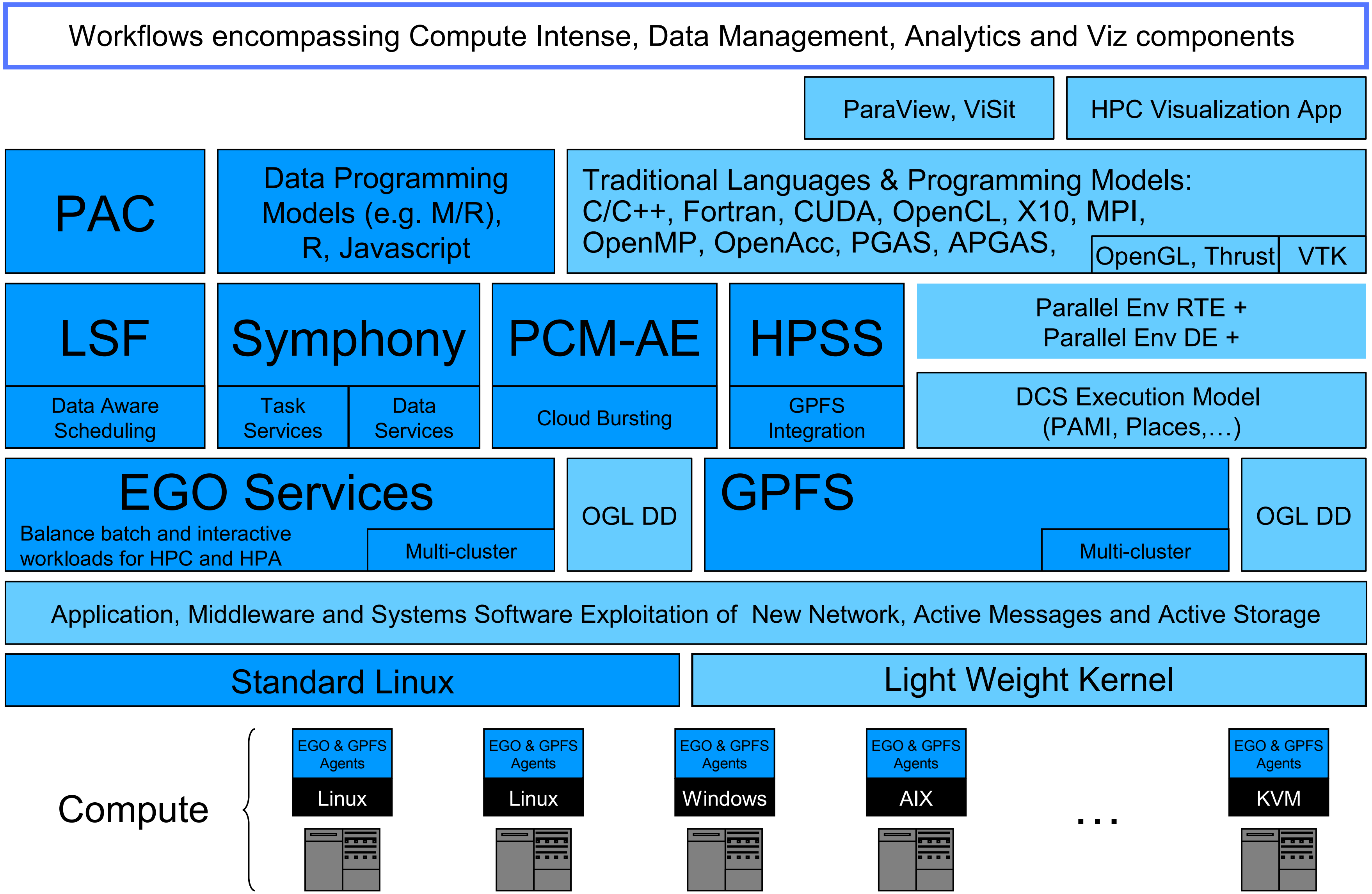
- Infrastructure management
  - Self-service cluster provisioning & management
  - Cluster flexing
- Self-service
  - Self-service submission & management
  - Dynamic provisioning, job migration & checkpoint-restart
  - 2D/3D remote visualization
- Integrated solution – hardware, software, services



Benefits

- Optimize utilization, maximize throughput
- Eliminate costly, inflexible silos and increase reliability
- Improve user and administrator productivity

DCS Integrated Software Stack: Commercial & HPC - 2017



Scale-out Server Solutions

http://public.dhe.ibm.com/common/ssi/ecm/en/cld00224usen/CLD00224USEN.PDF

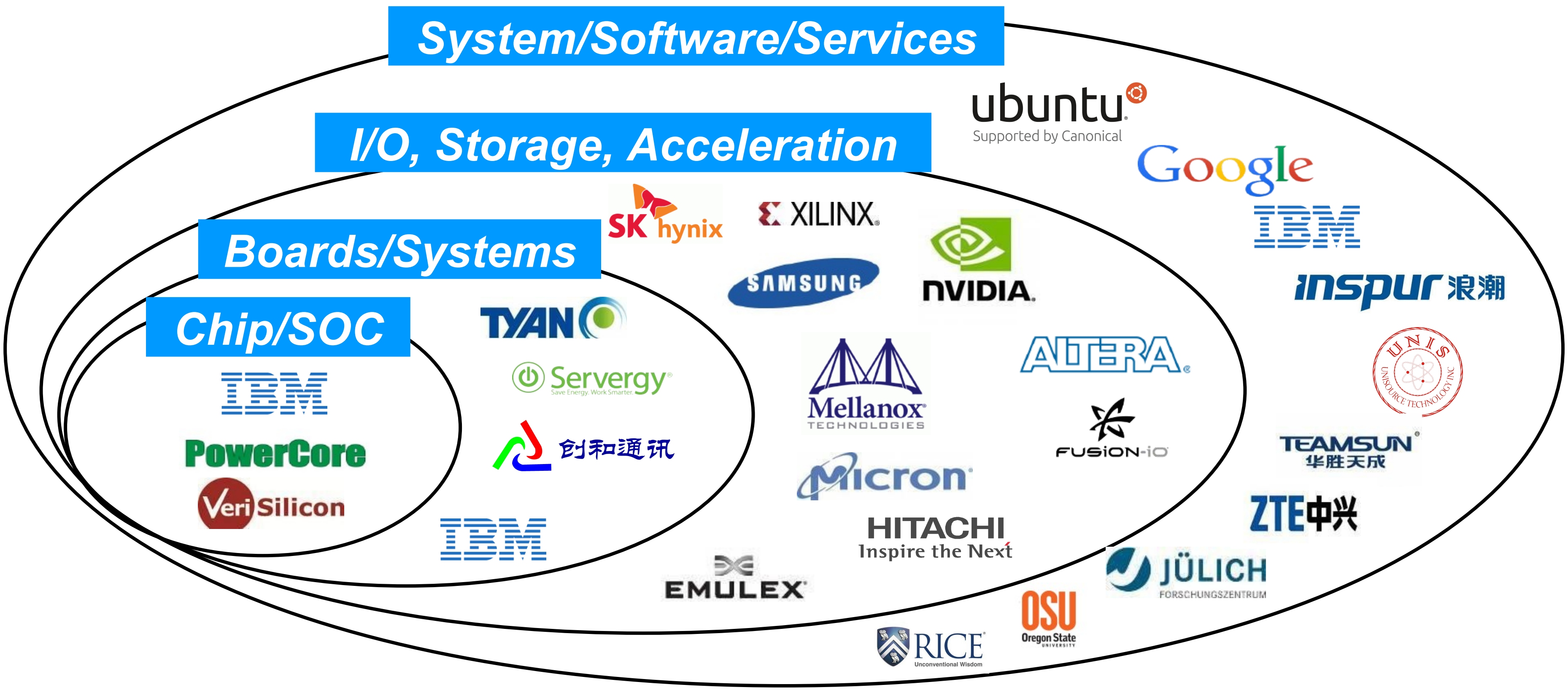
OpenPOWER Foundation was launched in 2013

Mission Statement:
The goal of the OpenPOWER Foundation is to create an open ecosystem, using the POWER Architecture to share expertise, investment, and server-class intellectual property to serve the evolving needs of customers. OpenPOWER Foundation is an open, not-for-profit technical membership organization that will enable today's data centers to rethink their approach to technology.

- Opening the architecture to give the industry the ability to innovate across the full Hardware and Software stack
  - Includes SOC design, Bus Specifications, Reference Designs, FW OS and Hypervisor Open Source
- Driving an expansion of enterprise class Hardware and Software stack for the data center
- Building a vibrant and mutually beneficial ecosystem for POWER

www.open-power.org

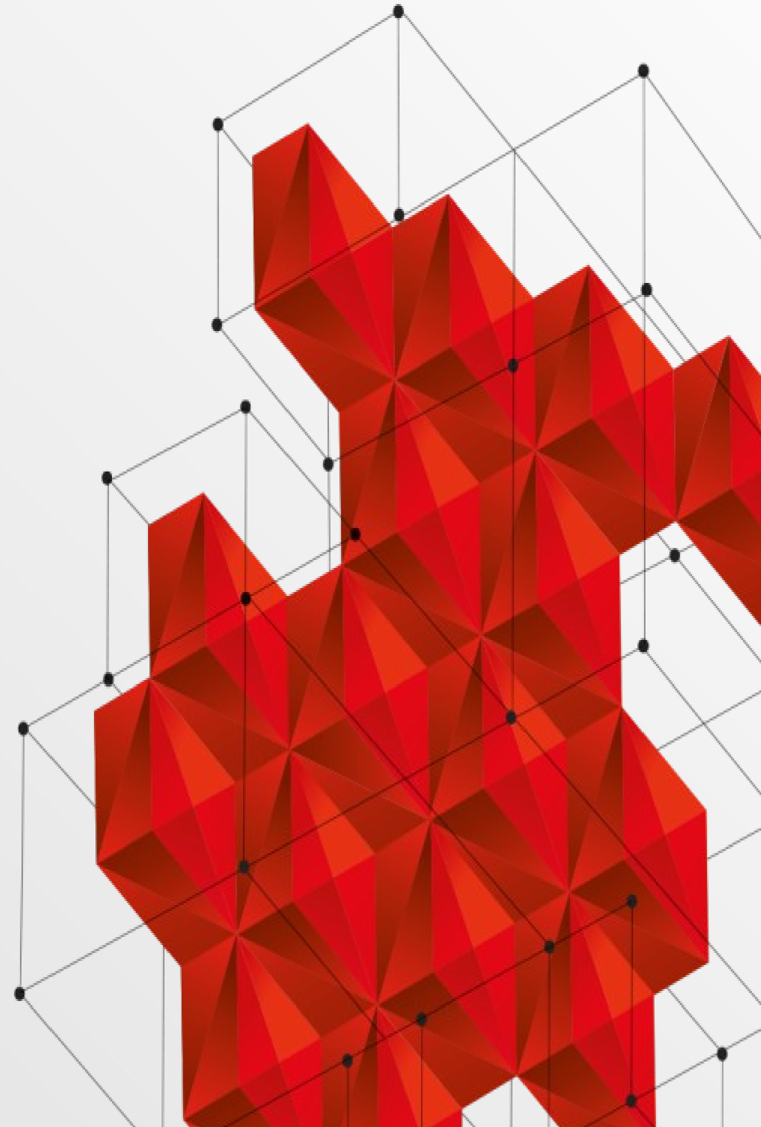
Membership growth confirms OpenPOWER opportunity



© OpenPOWER Foundation 2014

System x M5 Servers

Introducing IBM NeXtScale System M5



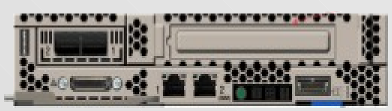
New Compute Node fits into existing NeXtScale infrastructure

One Architecture Optimized for Many Use Cases

Chassis
NeXtScale n1200 Enclosure

New

Compute node
IBM NeXtScale nx360 M5



- Dense Compute
- Top Performance
- Energy Efficient
- Air or Water Cool Technology
- Investment Protection

Storage NeX node*
nx360 M5 + Storage NeX

- Add RAID card + cable
- Dense 32TB in 1U
- Up to 8 x 3.5" HDDS

PCI Nex node (GPU / Phi)
nx360 M5 + PCI NeX

- Add PCI riser + GPUs
- 2 x 300W GPU in 1U
- Full x16 Gen3 connect

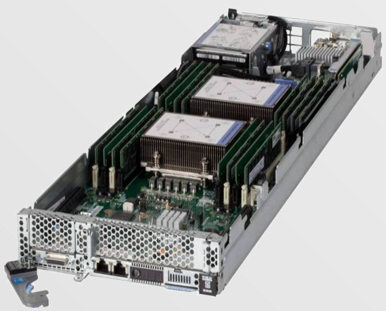
*M5 support to be available with 12Gb version at Refresh 1

NeXtScale - Choice of Air or Water Cooling

IBM NeXtScale System

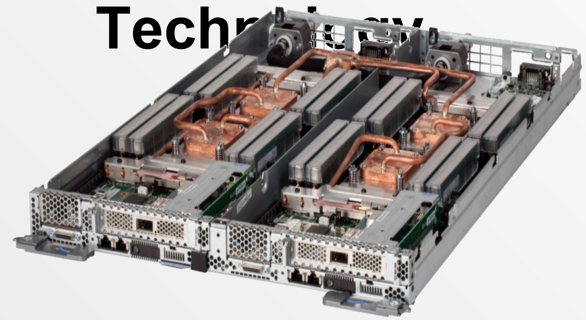
Your Choice

Air Cool



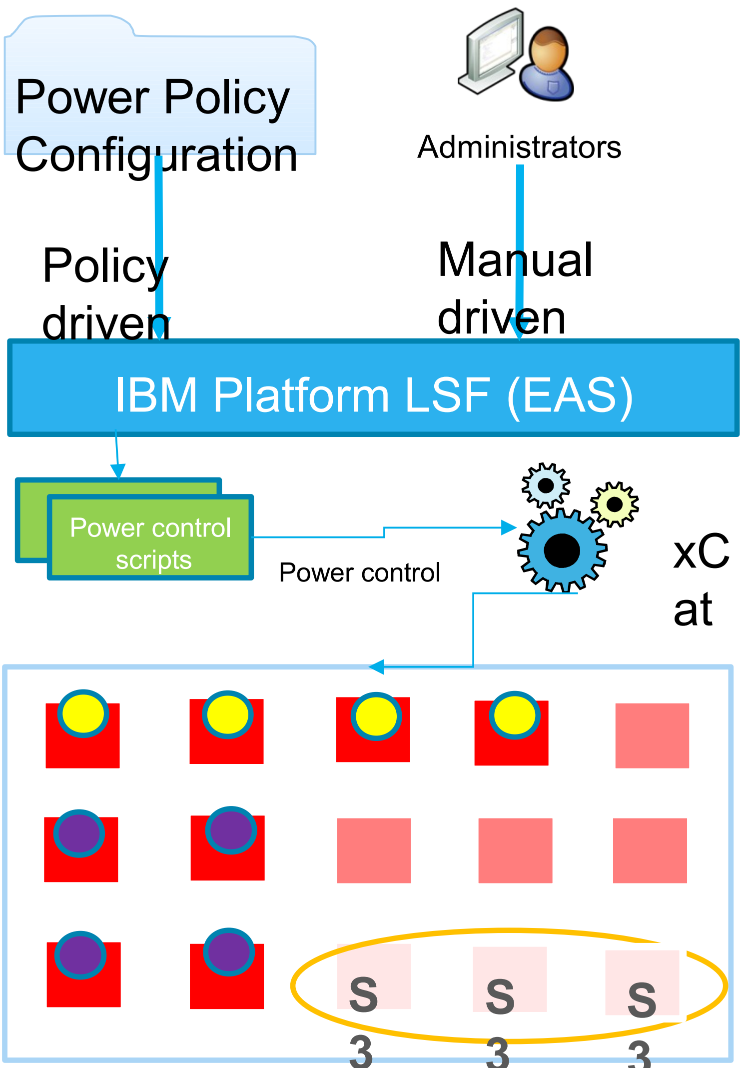
- Air cooled, internal fans
- Fits in any datacenter
- Maximum flexibility
- Broadest choice of configurable options supported
- Supports Native Expansion nodes
  - Storage NeX
  - PCI NeX (GPU, Phi)

Water Cool Technology



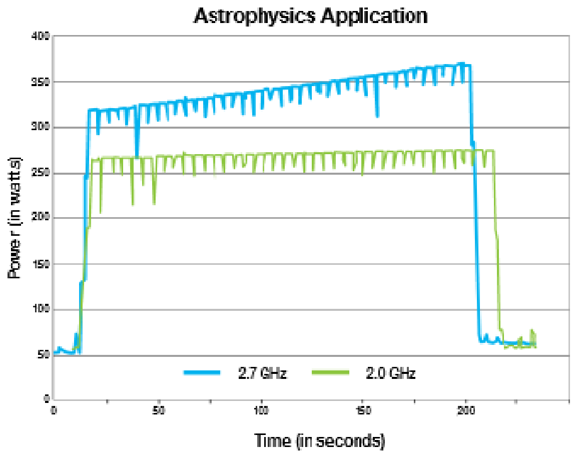
- Innovative direct water cooling
- No internal fans
- Extremely energy efficient
- Extremely quiet
- Lower power
- Dense, small footprint
- Lower operational cost and TCO
- Ideal for geographies with high electricity costs or space

How LSF supports Host Power Management?

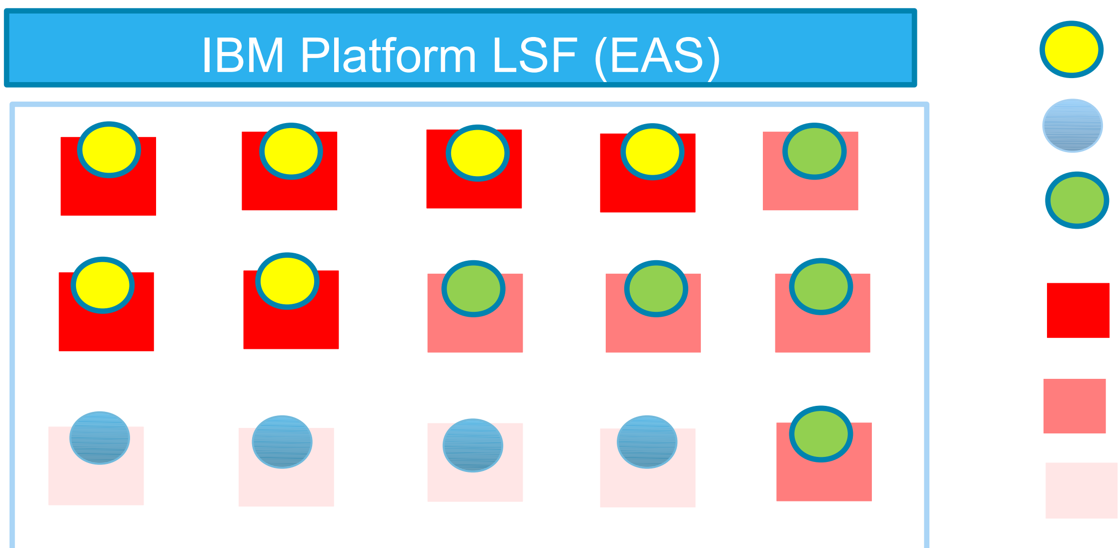


- **Manual management tools**
  - badmin hpower suspend/resume host
  - badmin hist
- **Policy Driven Power Saving**
  - Policy windows (i.e., 10:00 PM – 7:00 AM)
  - Node will be put in power saved (S3) if it is idle for a configurable period of time.
- **Power Saving Aware Scheduling**
  - Schedule jobs to use idle nodes first (Power saved nodes as last resort)
  - Aware of job request and wake up nodes precisely on demand
  - Safe period before running job on resumed nodes
- **Others**
  - customizable power control scripts
    - OOB xcat support
  - Power events tracking

Optimize Power Consumption of Active Nodes

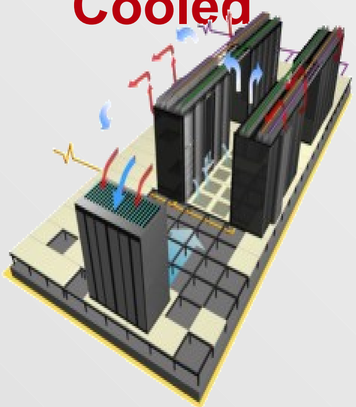


- Set a default cpu frequency on nodes
- Ability to set specified frequency on core/node level for a given job/application/queue
- Intelligently tag the job and select optimal cpu frequency based on energy and performance predication and site policies



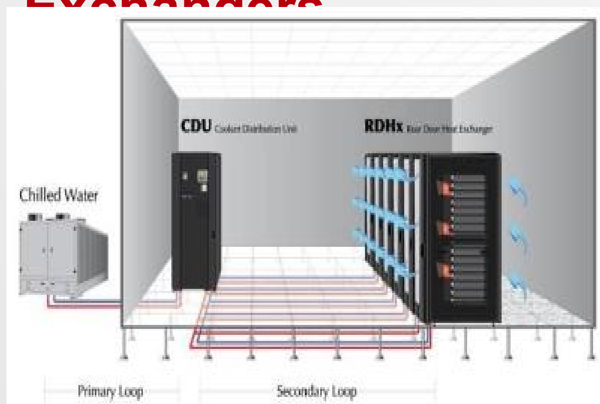
3 Ways to Cool your Datacenter

Air Cooled



- Standard air flow with internal fans
- Good for lower kW densities
- Less energy efficient
- Consumes more power – higher OPEX
- Typically used with raised floors which adds cost and limits airflow out of tiles
- Unpredictable cooling. Hot spots in one area, freezing in another

Rear Door Heat Exchangers



- Air cool, supplemented with RDHX door on rack
- Uses chilled water
- Works with all IBM servers and options
- Rack becomes thermally transparent to data center
- Enables extremely tight rack placement

Direct Water Cooled

- 100% water cooled
- No fans or moving parts in system
- Most energy efficient datacenter
- Most power efficient servers
- Lowest operational cost
- Quieter due to no fans
- Run processors in turbo mode for max performance
- Warm water cooling means no expensive chillers required
- Good for geographies with high electricity cost

Scale-out Storage solutions

http://public.dhe.ibm.com/common/ssi/ecm/en/cld00224usen/CLD00224USEN.PDF

Smart software-RAID technology without batteries

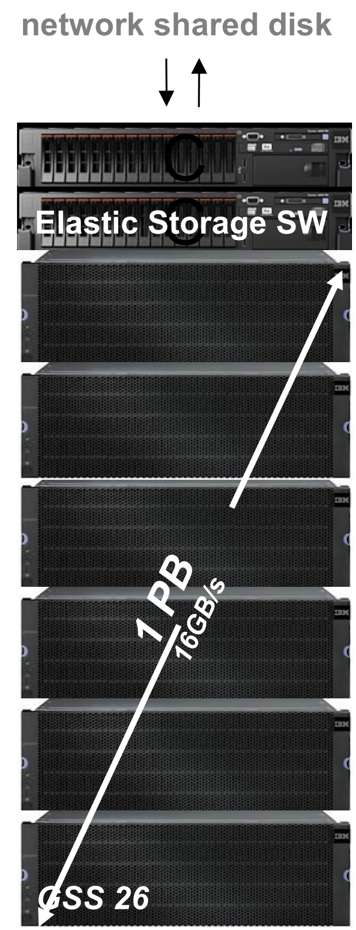
Elastic Storage Server

a hardened implementation of the Elastic storage cluster.

Currently available in two models with different flavors of nearline SAS drives and either InfiniBand or 10GbE. May be clustered with any existing GPFS 3.4+ cluster. May be clustered together to the max. node count supported by GPFS.

CRN Award 2013
Best Storage Innovation

Integrated.
Software-defined.
No NVRAM yet fast.
No batteries yet secure.
Massively scalable.

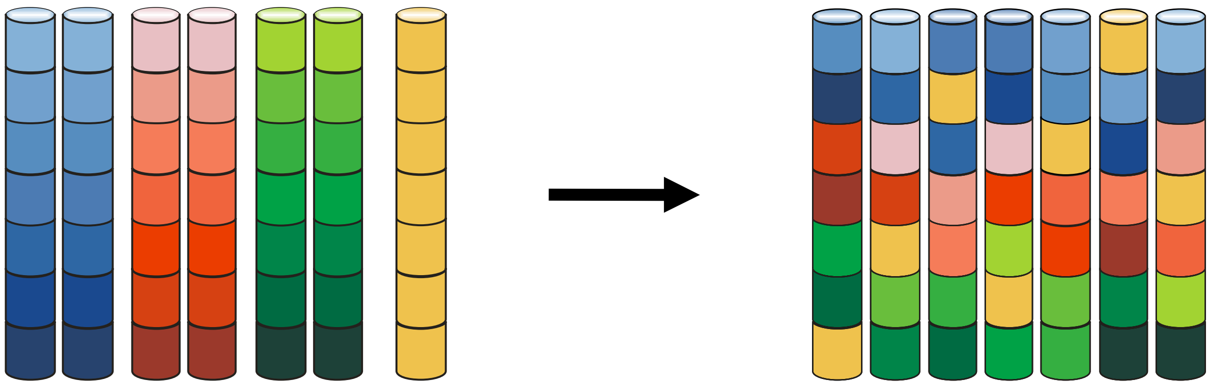


Elastic Storage Server

http://public.dhe.ibm.com/common/ssi/ecm/en/cld00224usen/CLD00224USEN.PDF

Declustered RAID Illustration

– Data, parity and spare strips are uniformly and independently distributed across disk array

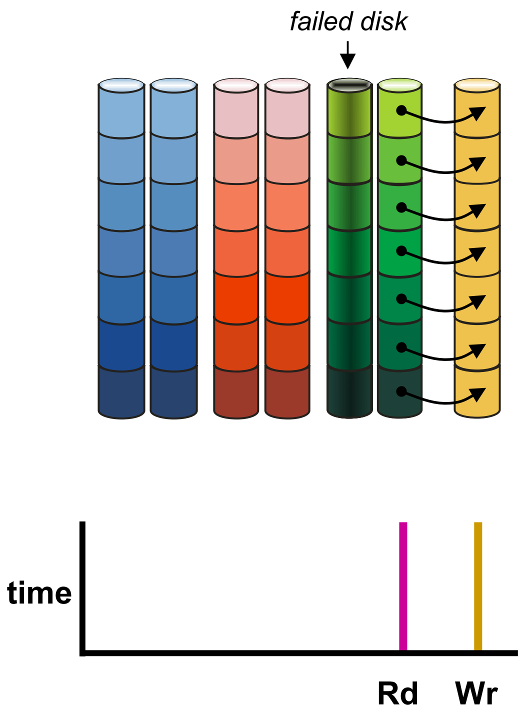


Conventional

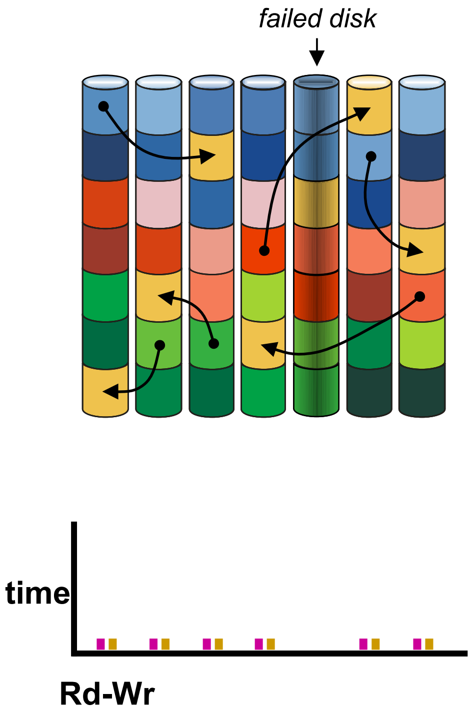
Declustered

– Supports an arbitrary number of disks per array
  - Not restricted to an integral number of RAID track widths

De-clustering can reduce data rebuild overhead by ~ 4-6 times



Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs

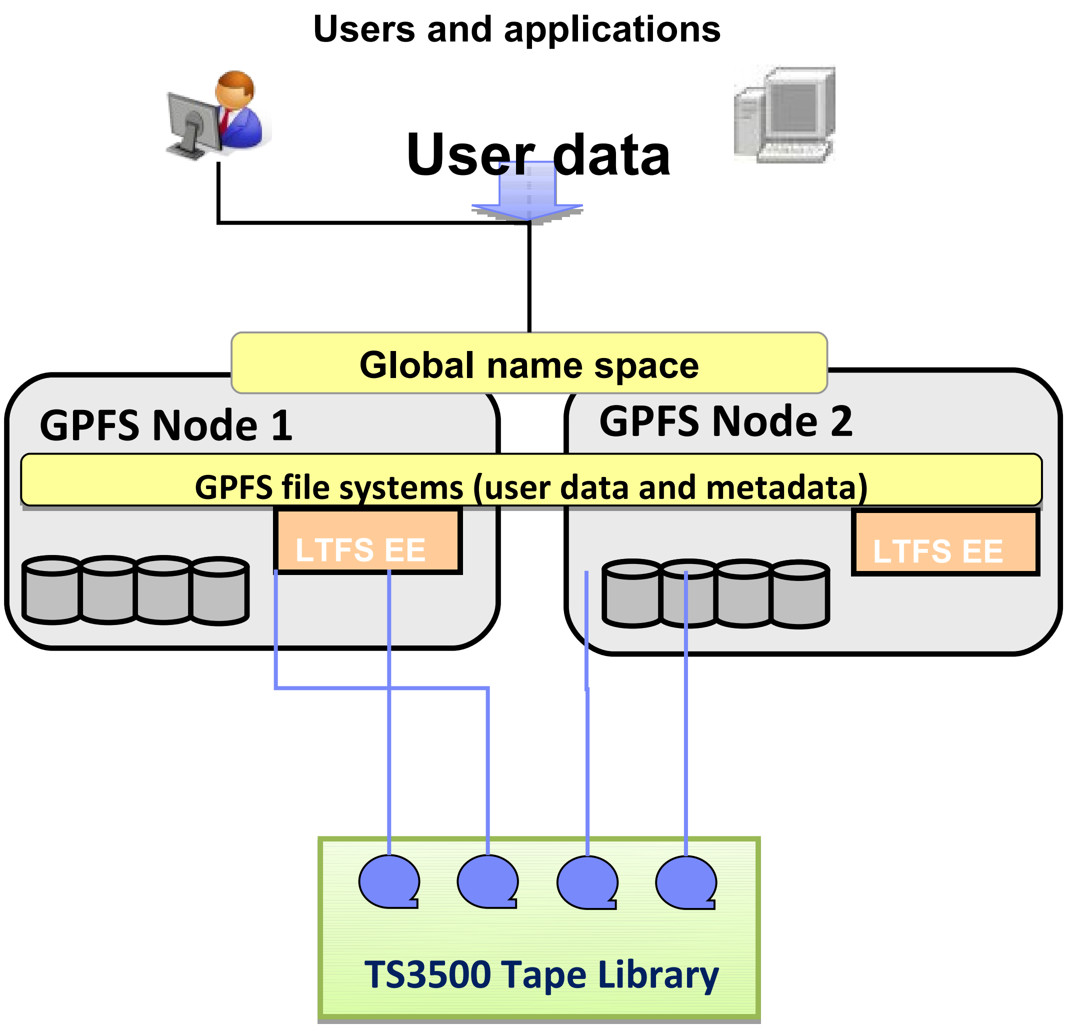


Rebuild activity spread across many disks, less disruption to user programs

Rebuild overhead reduced by 3.5x

LTFS EE system view

- **LTFS EE integrates LTFS LE with GPFS**
  - LTFS represents external tape pool to GPFS
  - Files can be migrated using GPFS policies or LTFS EE commands
- **GPFS provides global name space**
- **LTFS EE is installed on one or more GPFS nodes**
- **Workload is distributed over all LTFS EE nodes and tape drives**



The Solution – Tiered Network Storage

Operational

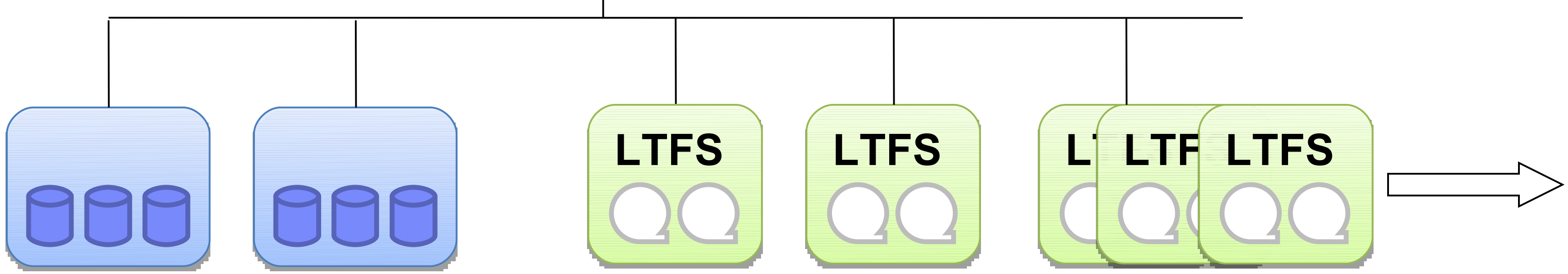
- Single file system view

C:/user defined namespace

High use data, databases, email, etc

Policy Based Tier Migration

Static data, rich media, unstructured, archive



Data Protection

Smaller

Scalable

- Easier to protect
  - Faster Time to recovery
  - Smaller backup footprint
- Time critical applications/data

- Lower cost, scalable storage
- Data types/uses for tape
  - Static data, rich media, etc.
- Replication backup strategies

TS3500 Tape Library Overview

A single library system includes:

- 1 - 16 frames (plus two service bay frames)
- 1 - 192 LTO and/or 3592 tape drives
- 59 - 20,000 storage slots
  - Up to 180 PB* Storage Capacity
- 16 - 224 I/O slots (1 - 14 I/O Stations)
  - 255 virtual I/O slots per logical library
- 1- 2 accessors (robotics)
- 0 - 15 shuttle connections to other library systems

Base frame (Model Lx3)

0 - 15 active expansion frames

0 - 15 shuttle connections

Service bay A

Service bay B

* Using TS1140 drives and media at 3:1 compression

Tape Drive Roadmap

	Gen 1 LTO1	Gen 2 LTO2	Gen 3 LTO3	Gen 4 LTO4	Gen 5 LTO5	Gen 6 LTO6	Gen 7 LTO7	Gen 8 LTO8
Native Capacity	100 GBs	200 GBs	400 GBs	800 GBs	1.5 TBs	2.5 TBs	6.4 TBs	12.8 TBs
Native Throughput (MBs/sec)	15	35	80	120	140	160	Up to 315	Up to 472
WORM	N/A	N/A	Yes	Yes	Yes	Yes	Yes	Yes
Encryption	N/A	N/A	N/A	Yes	Yes	Yes	Yes	Yes
Partitioning	N/A	N/A	N/A	N/A	Yes	Yes	Yes	Yes
Generally Available	Sept 2000	Feb 2003	March 2005	April 2007	April 2010	Nov 2012		

	Gen 1 3592	Gen 2 TS1120	Gen 3 TS1130	Gen 4 TS1140	Gen 5	Gen 6
Native Capacity	300 GBs	700 GBs	1 TB	4 TBs	8-10 TBs	14-20 TBs
Native Throughput (MBs/sec)	40	100	160	250	Up to 360	Up to 540
WORM	Yes	Yes	Yes	Yes	Yes	Yes
Encryption	N/A	Yes	Yes	Yes	Yes	Yes
Partitioning	N/A	N/A	N/A	Yes	Yes	Yes
Generally Available	Sept 2003	Oct 2005	Sept 2008	June 2011		

Roadmap is an estimate of the Program's current intent and is subject to change without notice.