

DE LA RECHERCHE À L'INDUSTRIE

cea


www.cea.fr

Lecture on Parallel Filesystems

Introduction

Jacques-Charles Lafoucriere

ENSIIE| 2018



DE LA RECHERCHE À L'INDUSTRIE


cea

Parallel File Systems Lecture Outline

- Data Management in Data/Computing Centres
- Distributed and Parallel Filesystems
- Lustre
- GPFS
- pNFS
- Fault Tolerance
- Parallel I/O

SFP 2018

Parallel Filesystems | PAGE 2



DE LA RECHERCHE À L'INDUSTRIE


cea

www.cea.fr

Data Management in Data/Computing Centers

Jacques-Charles Lafoucriere

ENSIIE| 2018



DE LA RECHERCHE À L'INDUSTRIE

cea

Outline

- The need
- Computing Center Architecture
- Hardware Technologies
- Software Technologies
- Management of large data volume

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 4



Storage Needs in Computing Centers

DE LA RECHERCHE À L'INDUSTRIE
What is Scientific Computing?

THE GRAND CHALLENGE EQUATIONS

$$\begin{aligned}
 & B_i A_i = E_i A_i + \rho_i \sum_j B_j A_j F_{ji} & \nabla_x \vec{E} = -\frac{\partial \vec{B}}{\partial t} & \vec{F} = m \vec{a} + \frac{dm}{dt} \vec{v} \\
 & dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV & \nabla \cdot \vec{D} = \rho & Z = \sum_j g_j e^{-E_j/kT} \\
 & F_j = \sum_{k=0}^{N-1} f_k e^{2\pi i j k/N} & \nabla^2 u = \frac{\partial u}{\partial t} & \nabla_x \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J} & P(t) = \frac{\sum_i W_i B_i(t) P_i}{\sum_i W_i B_i(t)} \\
 & & p_{n+1} = r p_n (1 - p_n) & \nabla \cdot \vec{B} = 0 & \\
 & -\frac{\hbar^2}{8\pi^2 m} \nabla^2 \Psi(r, t) + V \Psi(r, t) = -\frac{\hbar}{2\pi i} \frac{\partial \Psi(r, t)}{\partial t} & & -\nabla^2 u + \lambda u = f \\
 & \frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \nabla) \vec{u} = -\frac{1}{\rho} \nabla p + \gamma \nabla^2 \vec{u} + \frac{1}{\rho} \vec{F} & & \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f
 \end{aligned}$$

• NEWTON'S EQUATIONS • SCHROEDINGER EQUATION (TIME DEPENDENT) • NAVIER-STOKES EQUATION •
 • POISSON EQUATION • HEAT EQUATION • HELMHOLTZ EQUATION • DISCRETE FOURIER TRANSFORM •
 • MAXWELL'S EQUATIONS • PARTITION FUNCTION • POPULATION DYNAMICS •
 • COMBINED 1ST AND 2ND LAWS OF THERMODYNAMICS • RADIOSITY • RATIONAL B-SPLINE •

[Courtesy of San Diego Supercomputer Center]

SFP 2018
Parallel Filesystems / Computing Centers | PAGE 6



Scientific Computing

Why should we care about scientific computing?

- To complete experimental methods
- Computational simulations could be the only possible approach to analyze a problem:
 - Experiments may be cost prohibitive
 - Parametric study
 - Experiments may be impossible or forbidden



SFP 2018

Parallel Filesystems / Computing Centers | PAGE 7



Units and Order Of Magnitude

Units		Example
Byte		
Kilo Byte	KB = 10^3 B	30 KB = 1 text page
Mega Byte	MB = 10^6 B	5 MB = 1 music file
Giga Byte	GB = 10^9 B	1 GB = 2H film
Tera Byte	TB = 10^{12} B	1 TB = 6 millions of books (50 % BNF)
Peta Byte	PB = 10^{15} B	1 PB = DVD stack of Monpartnasse tower high
Exa Byte	EB = 10^{18} B	1 EB = 50 000 years of DVD-quality video
Zetta Byte	ZB = 10^{21} B	1 ZB = total of data moved in 2011
Yotta Byte	YB = 10^{24} B	1 YB = ?

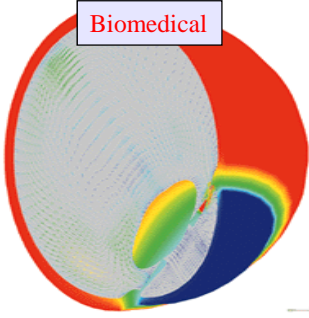
- Netflix uses few PB to store the video for streaming
- 5 PB of information produced up to 2003
- 466 EB shipped by hard drive industry in 2013
- 2.5 EB created by day in March 2015

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 8

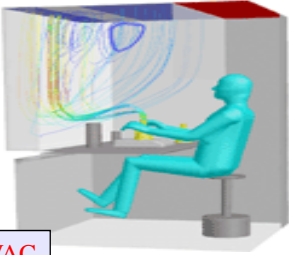
DE LA RECHERCHE À L'INDUSTRIE
cea **Examples of Scientific Computing**

Biomedical



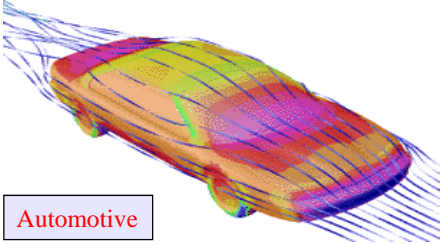
Temperature and natural convection currents in the eye following laser heating.

HVAC



Streamlines for workstation ventilation

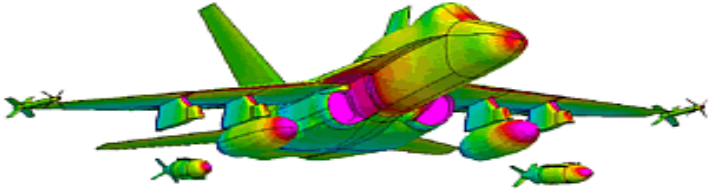
Automotive



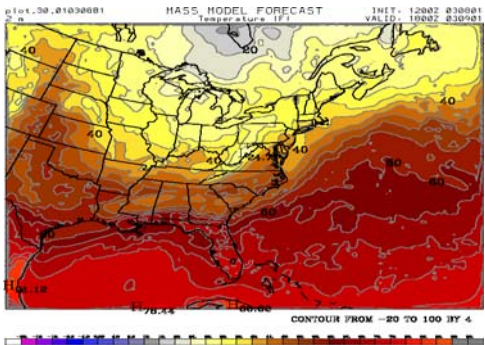
SFP 2018 Parallel Filesystems / Computing Centers | PAGE 9

DE LA RECHERCHE À L'INDUSTRIE
cea **Examples of Scientific Computing**

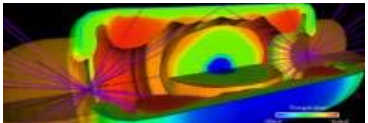
Aerospace



F18 Store Separation



Weather Forecasting



High-Energy Laser-Target Interactions


SFP 2018 Parallel Filesystems / Computing Centers | PAGE 10


DE LA RECHERCHE À L'INDUSTRIE

Examples of Scientific Computing

L'Oréal R&D
 Hair simulation

▶ Realistic movement simulation

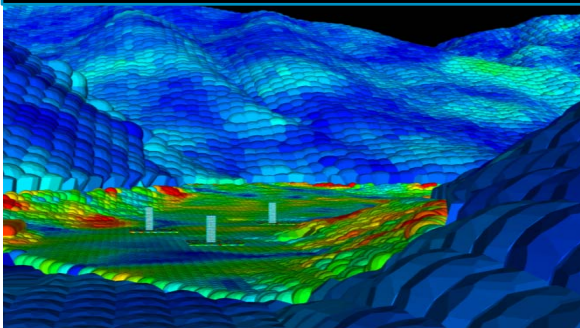




L'Oréal – Inria Collaboration

Sismology

▶ Risk evaluation
 ▶ Warning




SFP 2018
Parallel Filesystems / Computing Centers | PAGE 11

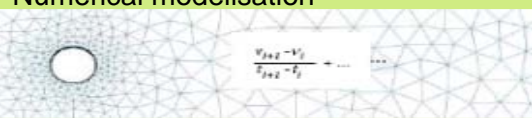
DE LA RECHERCHE À L'INDUSTRIE

Simulation process

Numerical solution




Numerical modelisation




Mathematical model

Physics equations

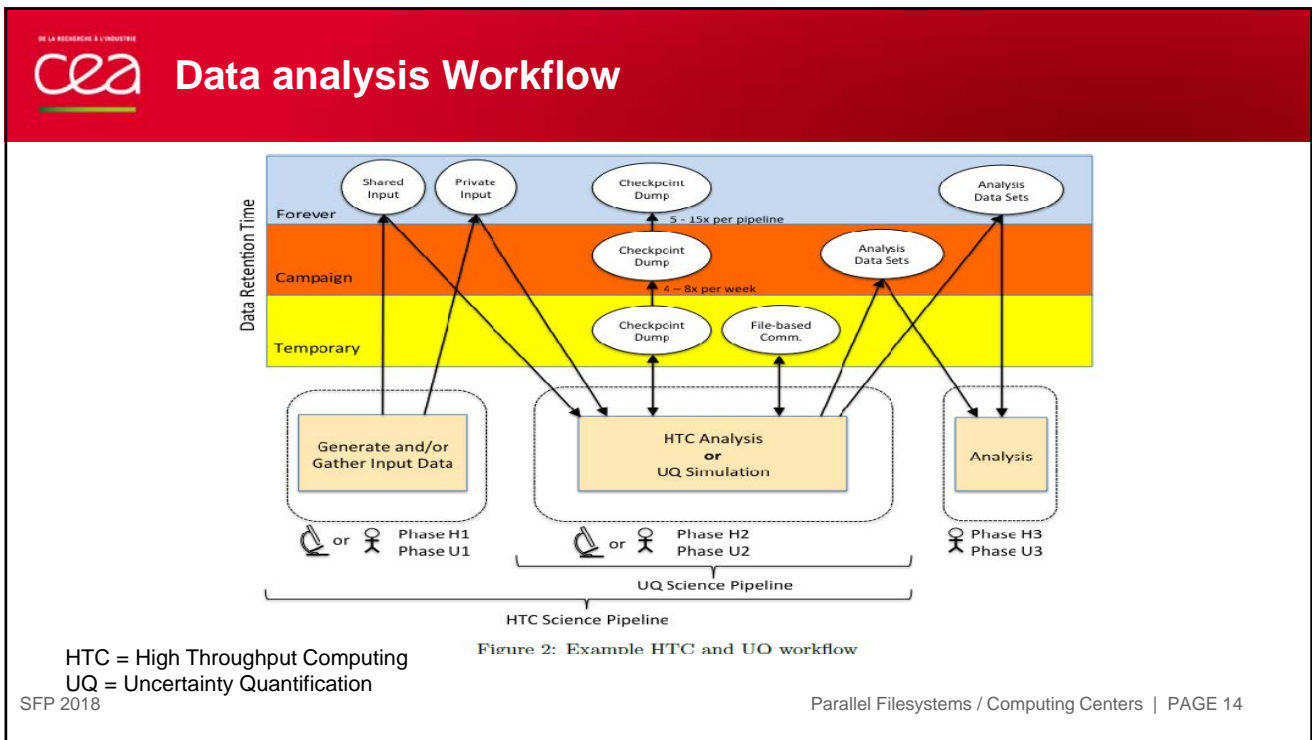
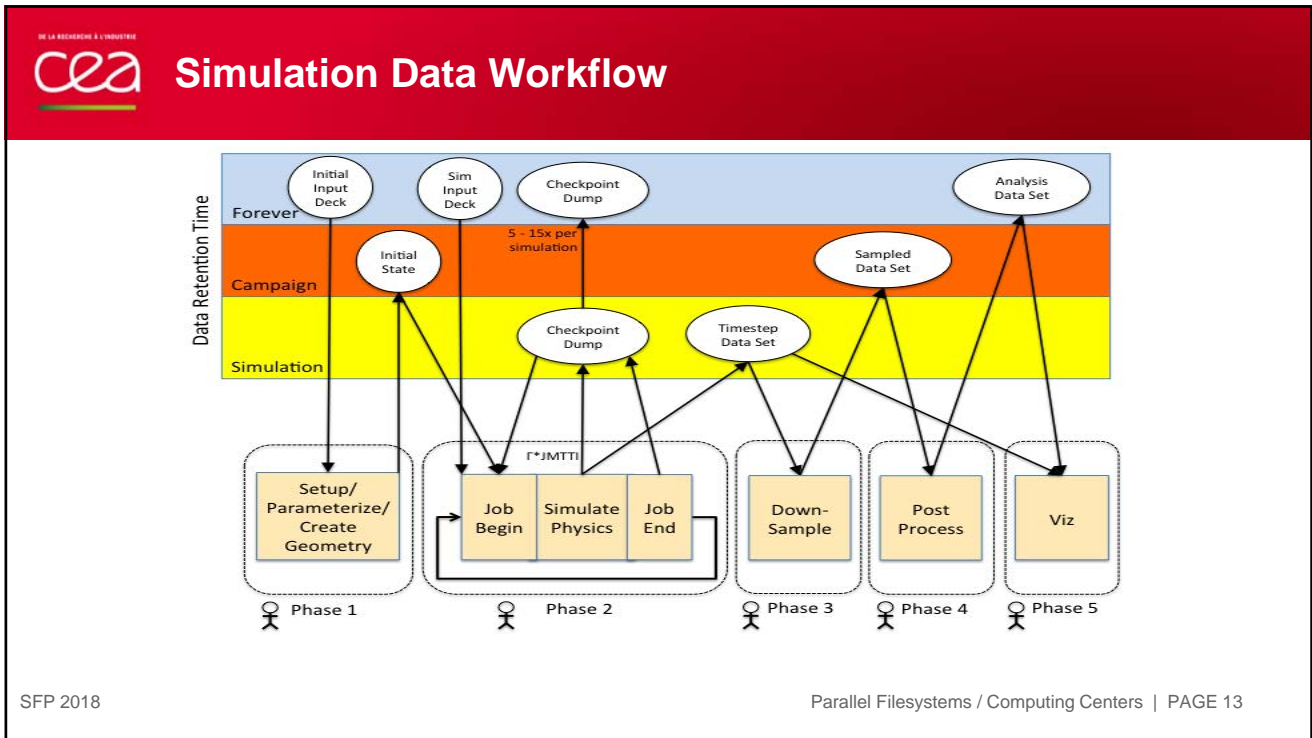


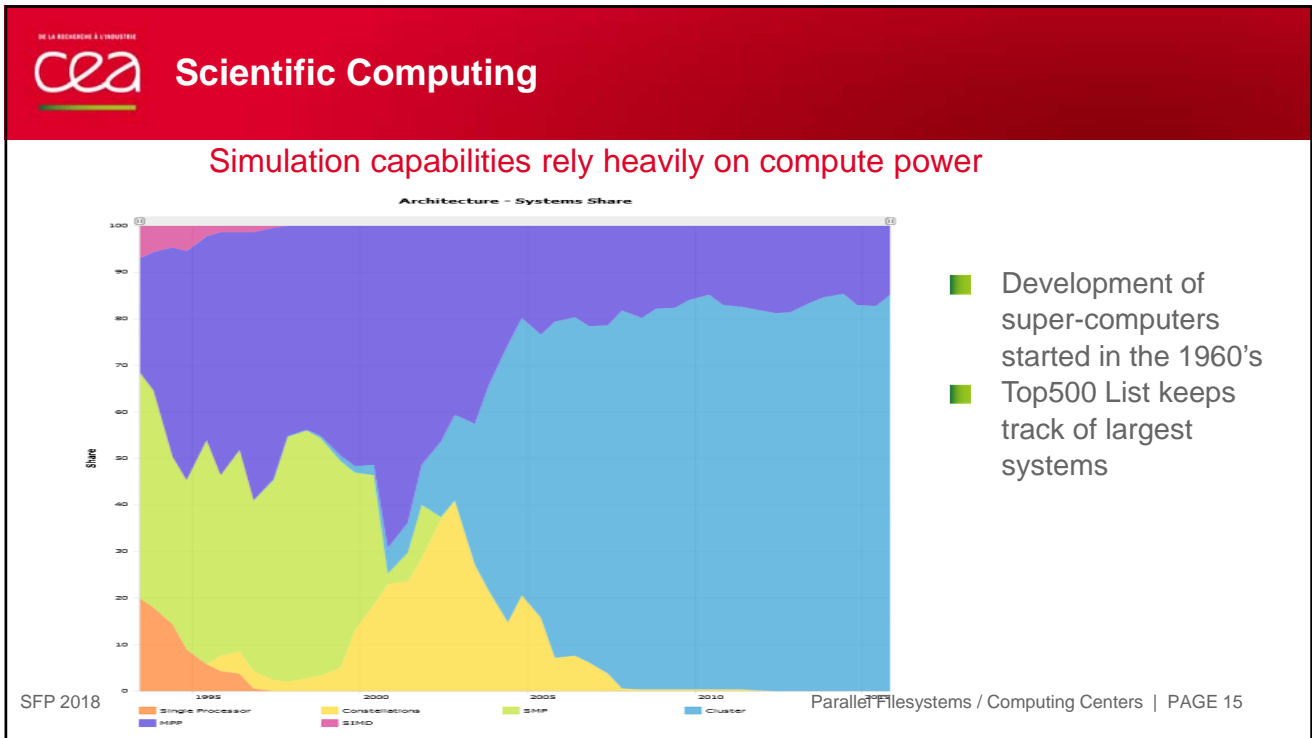
$$\rho \left(\frac{\partial v}{\partial t} + v \cdot \nabla v \right) = -\nabla p + \mu \nabla^2 v$$

Real problem



Parallel Filesystems / Computing Centers | PAGE 12





CEA: Example of Computing Center

| PAGE 16



Last Century

Compute Systems

- Few Cray Supercomputers (vectors and MPP)
- Few front-end machines



YMP

Storage Systems

- Directly connected to the front-end or to the super-computer
- Data managed through HSM



T90



T3E

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 17



Early 2000 and later

Cluster Age

- Tera 1
 - 5 Tflops
 - Disks: 50 TB, 7 GB/s
 - Tapes: 1 PB
- Tera 10
 - 60 Tflops
 - Disks: 2 PB, 100 GB/s
 - Tapes: 10 PB
- Curie or Tera 100
 - Over 1 Pflops
 - Disks: 20 PB, 500 GB/s
 - Tapes: 30 PB



TERA1



TERA10



TERA100




Curie

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 18

cea Today: dataless compute clusters

- Tera 1000
 - Phase 1
 - 2.6 Pflops
 - Phase 2
 - 30 Pflops
 - Disks: 40PB, 767 GB/s
 - Tapes: 80 PB

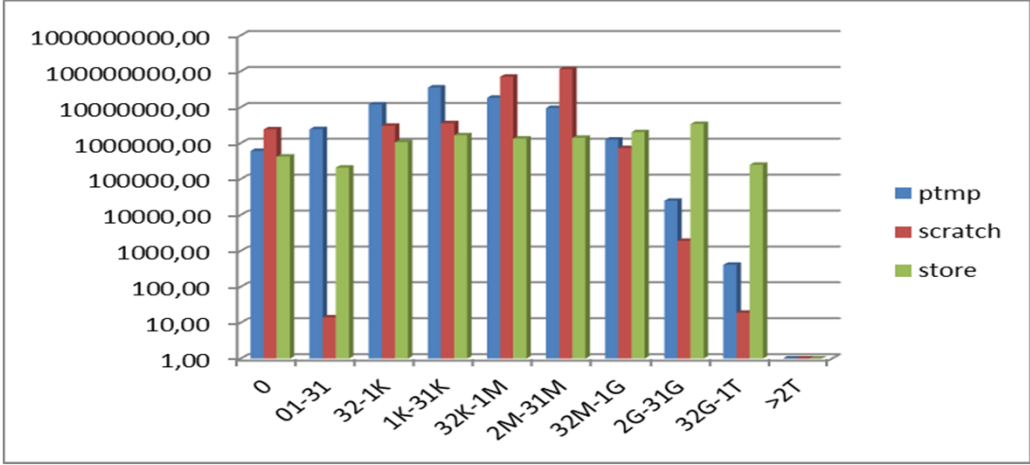


TERA1000

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 19

cea TERA100: Data Stored

■ Data produced by Simulations



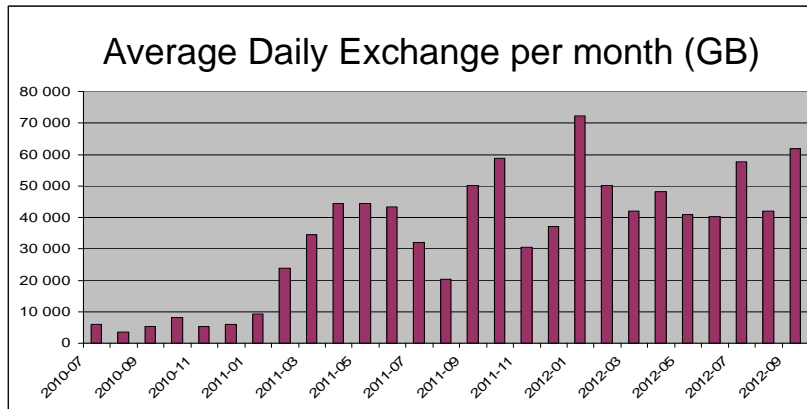
File Size	ptmp	scratch	store
0	~100,000,000	~1,000,000,000	~100,000,000
01-31	~100,000,000	~100,000,000	~100,000,000
32-1K	~100,000,000	~100,000,000	~100,000,000
1K-31K	~100,000,000	~100,000,000	~100,000,000
32K-1M	~100,000,000	~100,000,000	~100,000,000
2M-31M	~100,000,000	~100,000,000	~100,000,000
32M-1G	~100,000,000	~100,000,000	~100,000,000
2G-31G	~100,000,000	~100,000,000	~100,000,000
32G-1T	~100,000,000	~100,000,000	~100,000,000
>2T	~100,000,000	~100,000,000	~100,000,000

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 20

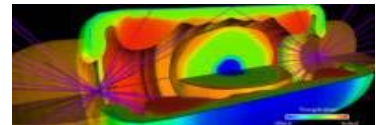


TERA100: Data Usage

Data moved between Compute Cluster and Storage Cluster



Numerical Simulations Data



Curie: Large Simulations

DEUS Grand Challenge

Une première mondiale réalisée en cosmologie sur CURIE : la modélisation de l'évolution de la structuration de tout l'Univers observable, du Big Bang à nos jours. Objectif : déterminer la nature de l'énergie noire qui constitue l'essentiel de l'Univers.

Jean-Michel Alimi est directeur de recherche première classe au CNRS. Il poursuit ses travaux de recherches au sein du Laboratoire Univers et Théories (LUTH), département de l'Observatoire de Paris, unité de recherche au CNRS associée à l'Université Paris-Diderot, qu'il a fondé le 1^{er} janvier 2002 et dirigé jusqu'au 31 décembre 2010. Il est l'un des grands spécialistes mondiaux de l'étude de l'origine, de l'évolution, de la structuration de l'Univers, et de la nature de la matière noire et de l'énergie noire. Initiateur du calcul massivement parallèle en astrophysique numérique au début des années 90, il favorise l'émergence d'une école française en formant plusieurs jeunes chercheurs.



76 000 cores
300 TB of memory

5 PB of data produced
Reduced to 500 TB

cea DEUS: a real use case

- 50 GB/s FS
- Close to max FS usage

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 23

cea New Needs: Experimental Data Analysis

Bioinformatics community

- New generation of DNA sequencer produces 4 TB / instrument / month
- Continuous data stream
- No data remove
- Data analysis produces 3 times initial volume

U.S. DOE Joint Genome Institute (JGI)
A 10-Year Strategic Vision (September, 2012)
Sequencing Output
(based on FY11 budget)

20 TB today +5 yrs +10 yrs 10 PB

Volume sur bandes évolutions futures

Zone tampon HSM Total

France Génomique Project
(to deployed at TGCC in 2013)

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 24



Data Deluge

Data are at the heart of computing centers

- Data management is a big challenge
 - For sys admin
 - For end users
- Data sets will grow
- Global data volume will grow
- Data use will increase

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 25



Data Found in Computing Centers

3 Classes of Storage

- HPC data
 - Large, Structured
 - POSIX API
- Cloud Based Storage
 - Static
 - REST Full API
- Big Data Analytics
 - Unstructured data

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 26

Computing Center Architecture

| PAGE 27

DE LA RECHERCHE A L'INDUSTRIE

Storage technologies

Media Characteristics

- Capacity
- Bandwidth
- Latency (access time)
- Reliability
- Size/Density
- Power Consumption
- Cost

Media Hierarchy

- Tapes
- Spinning Disks
 - Capacity Disks
 - Fast Disks
- Flash Memory

++ Capacity, --Speed, --E. Power

↓

-- Capacity, ++Speed, ++ E, Power

Cost

- No more capacity
- Mainly bandwidth
- Data movement is the new main cost

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 28

DE LA RECHERCHE À L'INDUSTRIE
cea **Media cost**


Cost	High Speed memory	RAM	Flash	HDD	Tape
BW \$/(GB/s)	10	10	300	2 000	30 000
Capacity \$/GB	?	8	0.3	0.05	0.01

2016

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 29


DE LA RECHERCHE À L'INDUSTRIE
cea **Media Aggregation**

Tape Robotics



Oracle Storagetek SL8500

Disks Controllers



DDN SFA10K

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 30

cea Data Management for Cluster

Data Centric Architecture

- Around a storage network
 - Dedicated to data movement
 - Independent of computing center backbone
- Hierarchical levels of storage
 - Fast storage (L1)
 - Large storage (L2)
- Initially based on the archival system: HPSS (T1 + T10)
 - T1: L1 = striped fast tapes
 - T10: L1 = striped disks
 - NFS access to HPSS name space (Ganesha)
 - CEA parallel tool to move data between HPSS and compute cluster
- Now based on parallel file system (T100/T1K + TGCC)
 - Lustre based
 - Lustre transparently migrated to HPSS

SFP 2018
Parallel Filesystems / Computing Centers | PAGE 31

cea File System Components

Petaflop File System

Compute Nodes ~4000 machines
Clients, POSIX access

I/O Servers
~200 machines
Data + Méta-Data

RAID Controllers ~20 000 disks

SFP 2018
Parallel Filesystems / Computing Centers | PAGE 32

Hardware Technologies



Hardware Components in a Cluster

Compute Node

- Xeon/Power/ARM servers
- High Performance CPU and memory
- As simple as possible to limit cost

Network

- High performance

File Servers

- Xeon/Power/ARM servers
- With a lot of I/O bandwidth

Storage Controllers

- Specialized hardware
- Manage/aggregate disks

DE LA RECHERCHE À L'INDUSTRIE
Network

Characteristics

- High bandwidth
 - How much data can go through the link
 - In GB/s
- Low latency
 - How fast a small (empty) message travel between to nodes
 - Few μ s
- Use Remote Direct Memory Access (RDMA)
 - Suppress memory copies
 - Reduce host CPU consumption

Examples

- InfiniBand
 - EDR (100 Gb/s)
- Ethernet
 - 10, 40, 100 Gb/s
- HPC
 - Cray Aries
 - Bull BXI
 - Intel Omni-path
- All based on links aggregation

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 35

DE LA RECHERCHE À L'INDUSTRIE
Network Topology

Main topologies

- Fat Tree
- 2D Torus
- 3D Torus
- Fully connected

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 36

DE LA RECHERCHE À L'INDUSTRIE
cea **Disk Controller**

RAID

- Was: Redundant Array of Inexpensive Disks
- Now: Redundant Array of Independent Disks



Multiple disk drives working together to

- Increase capacity of a single logical volume
- Increase performance
- Improve reliability/add fault tolerance




LSI Pikes Peak
Parallel Filesystems / Computing Centers | PAGE 37

SFP 2018

DE LA RECHERCHE À L'INDUSTRIE
cea **RAID Levels**

- RAID 0: Striping
- RAID 1: Mirroring
- RAID 2: Striping with parity
- RAID 3: Striping with parity (bit interleaved)
- RAID 4: Striping with parity (synchronous block interleaved)
- RAID 5: Striping with parity (independent block interleaved, distributed parity)
- RAID 6: Generalization of RAID 5 with P blocks of parity for N blocks of data
 - Based on Galois Field Theory and Reed-Solomon coding
 - Today currently used with P = 2 and N = 8
- Parity declustering
 - Disks are grouped in a large pool
 - Parity blocks are spread over all disks
 - Give a better scalability in rebuild phases (so always with large configuration)
 - Also called Erasure Coding

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 38



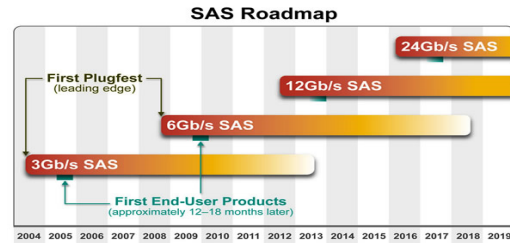
Storage Controller Access

Block mode access

- Storage is seen as few large continuous spaces read/write by blocks of few KB, up to MB

Storage Protocols

- SCSI (ancestor, parallel bus)
- Fibre Channel
 - Design for large storage fabrics
 - Partitioning, long distance
- SAS
 - Design for servers
- SATA
 - Design for desktop
- Aggregated serial links
- After years or standard wars, strong tendency to use shared technologies like cables



SFP 2018

Parallel Filesystems / Computing Centers | PAGE 39

Software Technologies



Parallel Filesystems

Use multiple servers together to aggregate disks

- Single name space from distributed nodes
- Improved performance
- Even higher capacities
- May use high-performance network

Vendors/Products

- Lustre (Intel)
- GPFS (IBM)
- GFS (RedHat)
- Gluster (RedHat)
- pNFS (EMC, NetAPP, IBM)
- Ceph (Inktank/RedHat)

SFP 2018

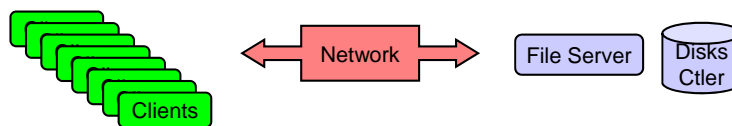
Parallel Filesystems / Computing Centers | PAGE 41



Network based: NFS

Basic Client/server mode

- Design ~80
- Standard
- Based on RPC over tcp/ip
- No coherency between clients
- Single server, not scalable
- Good for login home, and system configuration, tools



SFP 2018

Parallel Filesystems / Computing Centers | PAGE 42

cea Network based: pNFS

Parallel Client/server mode

- NFS evolution (NFS v4.1)
- Standard
- Based on RPC over tcp/ip and RDMA
- No coherency between clients
- 1 server manages file layouts
- Multiple servers manage data
 - Multiple data access models are supported (file, block, object)
- Client data access is parallel

The diagram illustrates the pNFS architecture. On the left, a stack of green boxes represents multiple clients. A red double-headed arrow labeled 'Network' connects these clients to a central server architecture. This architecture includes an 'MD Server' (Metadata Server) and several 'Data Servers'. Each 'Data Server' is connected to its own 'Disks Ctlr' (Disk Controller). The MD Server manages the file layouts, while the Data Servers handle the actual data access in parallel.

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 43

cea SAN based: GPFS

Based on a shared storage

- Use a storage network to give controller storage access to all nodes
- Clients implement all the FS logic (// access)
- Clients manage locks and use a distributed lock manager (DLM) to warranty coherency
- Byte range locking
- To workaround SAN scalability issue, GPFS implements a software storage server
- MetaData and Data location can be optimized
- Scalable fault tolerance

The diagram illustrates the GPFS architecture in two parts. On the left, a stack of green boxes represents clients connected via a 'SAN' (Storage Area Network) to a shared storage controller labeled 'Disks Ctlr'. A red double-headed arrow labeled 'Model' is positioned above the SAN connection. On the right, a stack of green boxes represents clients connected via a 'Network' to a 'File Server' and a 'Disks Ctlr'. A red double-headed arrow labeled 'GPFS' is positioned above the network connection. This represents the software storage server approach where the File Server manages the data layout and the Disks Ctlr handles the physical storage access.

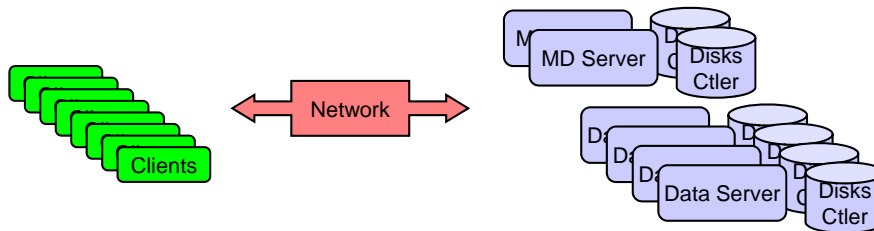
SFP 2018 Parallel Filesystems / Computing Centers | PAGE 44



Network based: Lustre

Based on MetaData and Data servers

- Servers are dedicated to MetaData and Data
- Clients implement a minimum of knowledge
- Clients first ask layout to MD server and after make // access to data servers
- Servers manage locking for their objects to warranty coherency (scalable)
- Byte range locking
- 2 servers fault tolerance model is limited
- Used on the largest systems of TOP500



SFP 2018

Parallel Filesystems / Computing Centers | PAGE 45



Special Features

Networking (Lustre only)

- Lustre network is based on an abstraction layer: LNET
- Based on RDMA model
- Support many networks (IB, Qsnet, Myrinet, ...)
- Support LNET routers
 - Allow building data center global FS
 - Isolate client from shared resources

Data Management

- Fileset and pools
- HSM binding
- Quotas
- QoS

SFP 2018

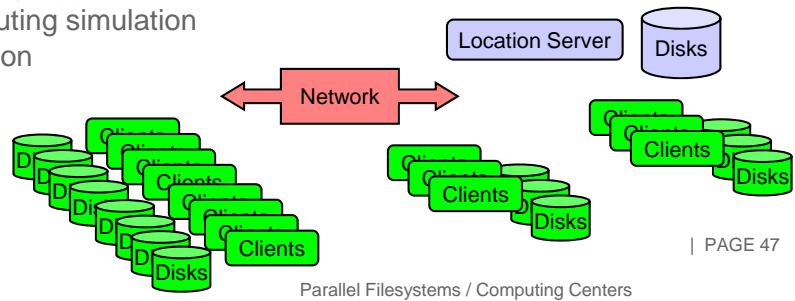
Parallel Filesystems / Computing Centers | PAGE 46



Hadoop

Avoid Data Movement

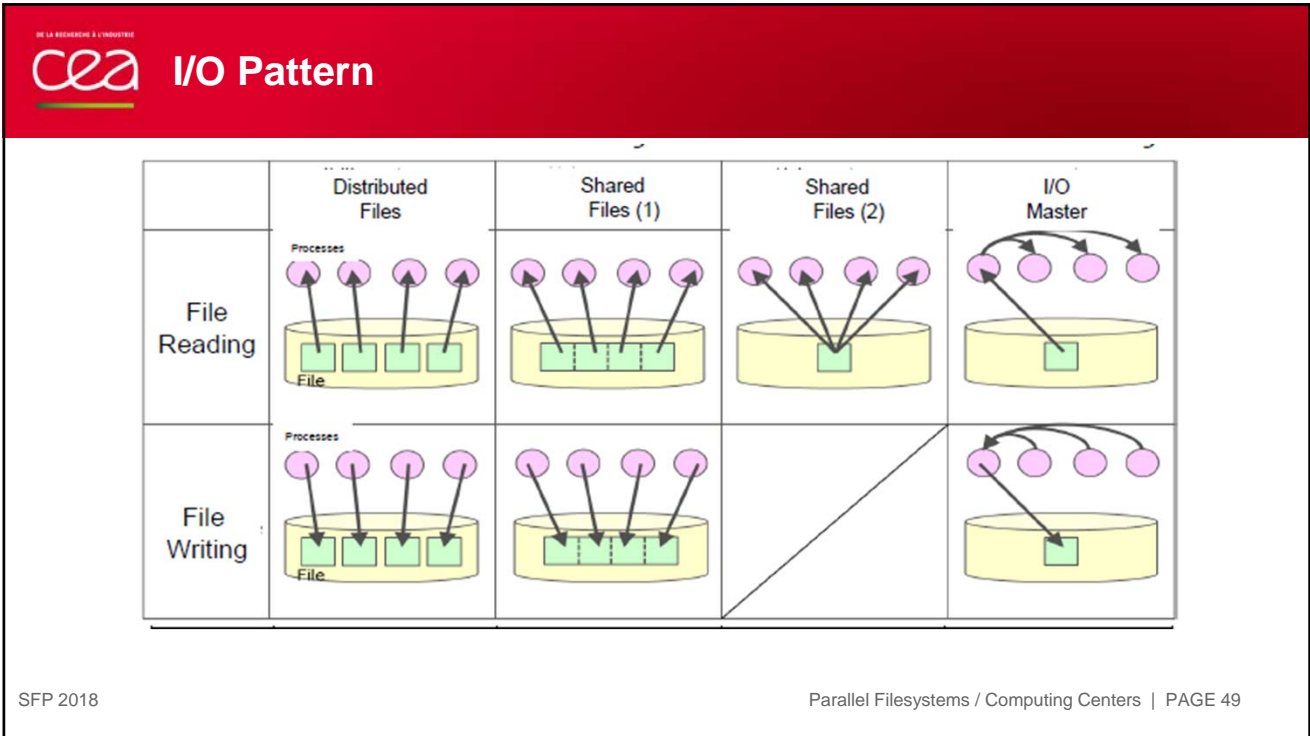
- Computation comes where is the data
- Use of locale storage only (cheap)
- Redundancy based on duplication
 - Use of physical location criteria to minimize impact
- Granularity is a file
 - No random modification, only append
 - Strong constraint for computing simulation
- Only one server manage file location
- Good for data analysis



| PAGE 47

SFP 2018

How to Access Data?



Management of large data volume



Challenges for large configurations

Scaling

- Standard 1-1 fail-over model does not scale well
 - 4-1 model: 3 servers can take the load of a failed one
- Lock management by clients does not scale well on heavy load
- Global efficiency and wall time resolution

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 51



Challenges for large configurations

Reliability

- Failure detection is based on timeout
 - Impossible to differentiate a failed node from a loaded node
- Large systems have always failed parts
 - Rebuild load must be under control
- Backup
 - Full backup is too long on a Peta sized file system : Impossible to make standard backups
 - Move to an event based model
 - All changes are registered in a SQL DB
 - File search based on SQL requests

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 52



Challenges for large configurations

User control

- A crazy node can easily killed any parallel file system
 - e.g.: Loop on a failed syscall
- No easy way to limit user use of bandwidth storage

Memory consumption

- Any file system need I/O buffers
- Be careful with memory usage linear with resources count (statistics, buffer reservation, ...)

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 53



Storage Performance Reproducibility

Storage Use

- Simulations do not use storage continuously but in burst mode
- Cluster storage resources are shared by runs
 - Too expensive to dedicate storage resources to run
- Different codes can be some time synchronous
 - Bandwidth is shared => difficult to have always the same performance
- Data placement is defined at file write
 - Read can only follow it
- Different solutions are under investigation
 - On demand bandwidth reservation
 - QoS in file systems
 - Node bandwidth limitation

SFP 2018

Parallel Filesystems / Computing Centers | PAGE 54

DE LA RECHERCHE À L'INDUSTRIE
cea Next?

Distributed File System and Parallel File system

SFP 2018 Parallel Filesystems / Computing Centers | PAGE 55

Thank you for your attention

ENSIIE | 2018

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM-Ile de France | 91297 Bruyères-le-Châtel Cedex
T. +33 (0)1 69 26 40 00 | F. +33 (0)1 69 26 70 86

Direction des applications militaires
Département sciences de la simulation et de l'information
Service informatique scientifique et réseaux

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019