

DE LA RECHERCHE À L'INDUSTRIE

**cea**

www.cea.fr

# Lecture on Parallel Filesystems

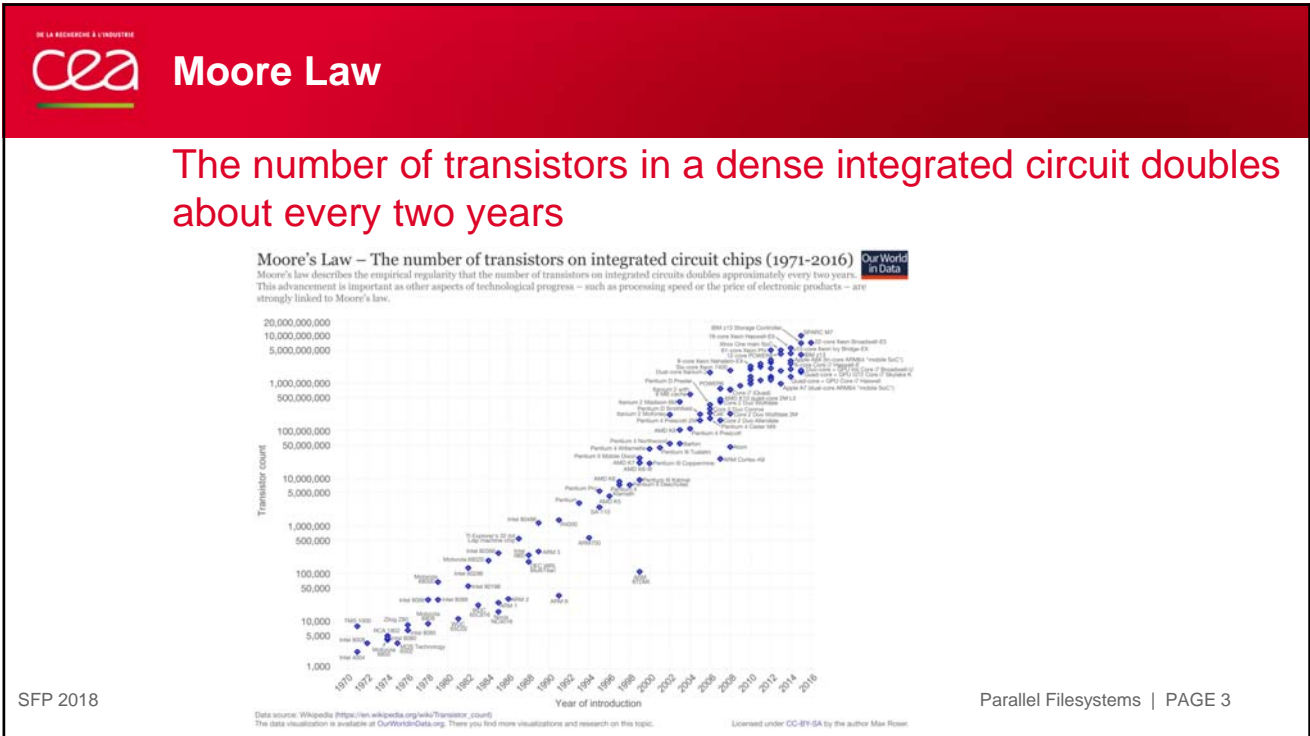
Outlook 2020

Jacques-Charles Lafoucriere

ENSIE| 2018

# ExaScale Challenges

| PAGE 2



DE LA RECHERCHE À L'INDUSTRIE

**cea** Energy Efficiency Challenge

Moore's law drives current systems to ~150 MW in 2020 for an exaflop supercomputer

- “Acceptable” infrastructure and running costs is 10 to 15MW => Target is ~1/10
  - Need components (processors, memories, I/O links,...) optimized for performance per Watt ratio
  - Need minimizing power distribution and cooling costs

These technological breakthrough will deeply impact

- System architecture (gigantic # of cores)
- Programming models & algorithm
  - Multi-scale parallelism (message passing, threads, vectors)
  - Compute vs data movement
  - ...

SFP 2018

Parallel Filesystems | PAGE 4

**cea** Heat Wall

Limit is processor cooling =>  $P = cte$

$$P_d = C_e \times F \times V^2$$

**Before**

- Reduce voltage
- Increase frequency

**Now need to do more computation with same power consumption**

- Increase efficiency
- Increase parallelism

SFP 2018 Parallel Filesystems | PAGE 5

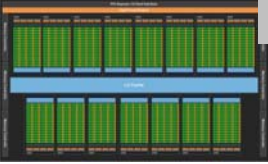
**cea** What Kind of Architecture?

**GPU and MIC deliver a significant boost to computation**

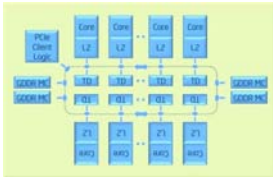
- But they are seen as co-processors
- Would be great to have standalone chips
  - No costly transfers needed
  - Simpler system administration
  - Vendors roadmaps are not stabilized yet

**Potential architectures**

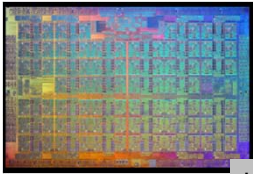
- Manycore will dominate
  - IF the codes are ready
  - IF the chips are up to the expectations
  - IF we can design a balanced machine
    - I/O will be the next bottleneck
    - The memory wall might be still there



Nvidia Kepler



Intel Phi



Intel KNL

SFP 2018 Parallel Filesystems PAGE 6



## How To Use It?

### Many levels of parallelism

- FPU
  - Vector instructions
  - Vector of 256 or 512 bytes
- CPU
  - Cores
  - Hyper threads
  - Ten's of compute threads
- Node
  - Multi CPU
  - Numa architecture
  - 2 or 3 levels of memory
- Cluster
  - Ten's of thousands of nodes

SFP 2018

Parallel Filesystems | PAGE 7



## How to Use it (Cont.)?

### High level of hybrid parallelism

- Vector instructions
- Multi threading (OpenMP)
- Message Passing (MPI)

### 3 levels of parallelism

- Micro (vectors)
- Meso (threads)
- Macro (MPI)

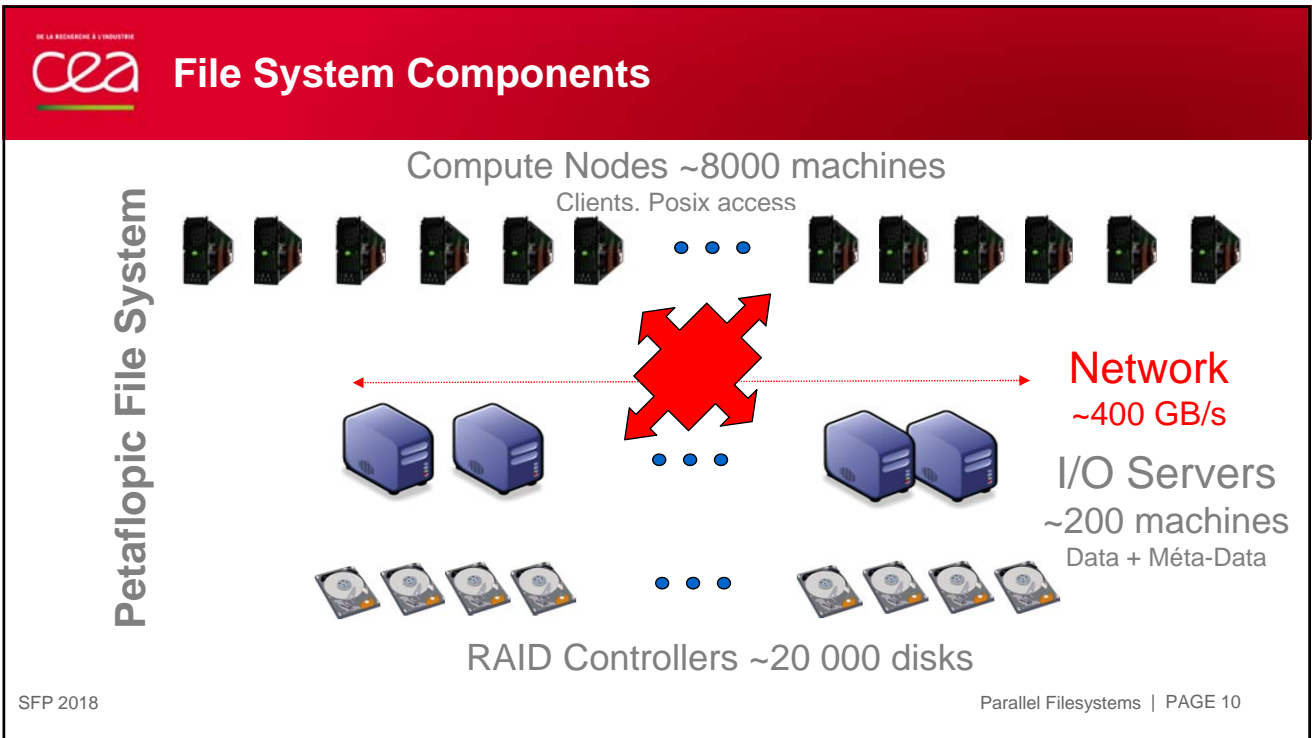
Each application has to choose the right balance for it's algorithms

SFP 2018

Parallel Filesystems | PAGE 8

# Computing Center Today

| PAGE 9





# Petaflopic Computing Centers



Tera  
30 Pflop/s  
Mem: 1 526 TB  
FS: 600 GB/s

TGCC  
10 Pflop/s  
Mem: 500 TB  
FS: 500 GB/s



SFP 2018

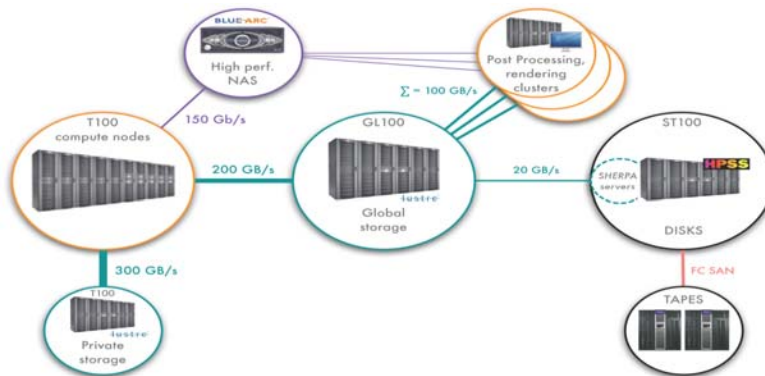
Parallel Filesystems | PAGE 11



# Data Centric Architecture

## TERA

- Design for heterogeneity
- Ready to integrate new needs

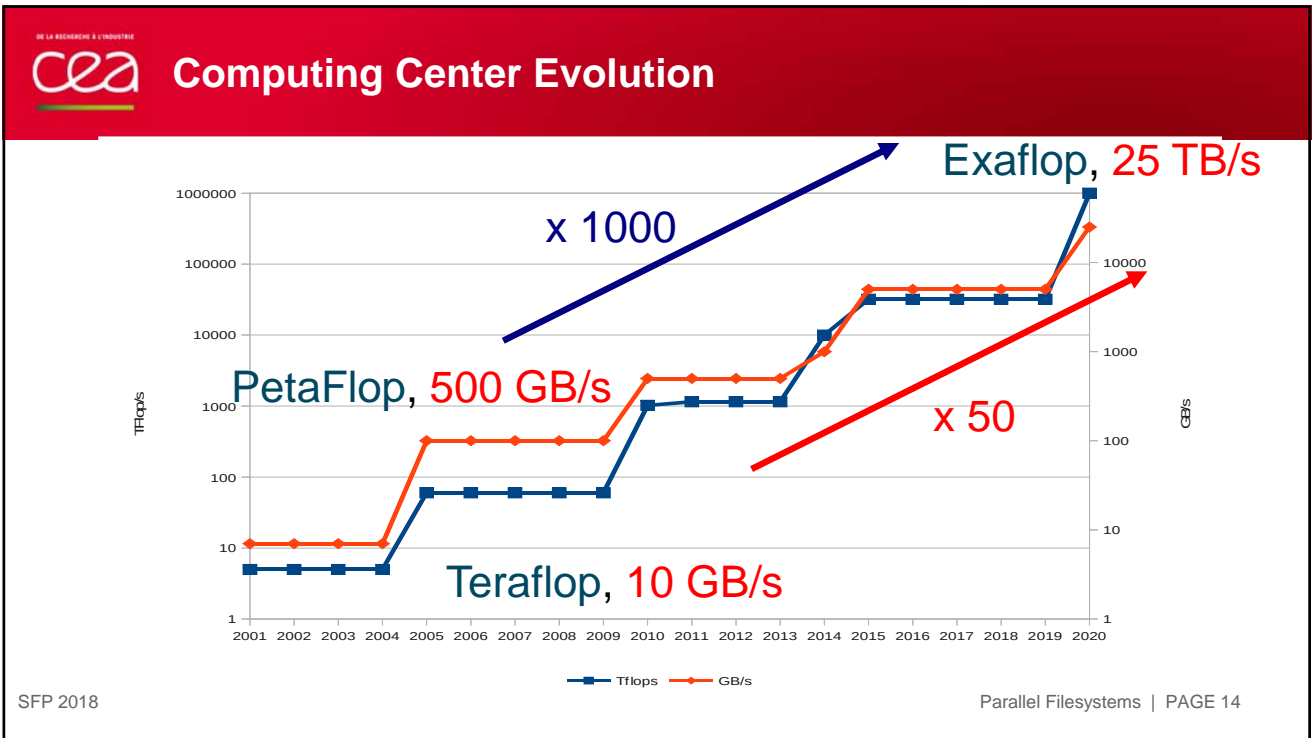


SFP 2018

Parallel Filesystems | PAGE 12

# Storage Evolution

| PAGE 13



DE LA RECHERCHE A L'INDUSTRIE  
**cea** **Roadmap to Get an Exaflop Class System**

2018	2020
<ul style="list-style-type: none"><li>■ 30 Petaflops</li><li>■ 3.5 To/s</li><li>■ 10 000 nodes</li><li>■ &gt; 20 000 disks</li></ul>	<ul style="list-style-type: none"><li>■ ~1 Exaflops</li><li>■ 25 To/s</li><li>■ &gt; 50 000 nodes</li><li>■ &gt; 30 000 disks</li></ul>

**Impossible to keep the same ratio for storage bandwidth and for compute power**

**New type of compute nodes**

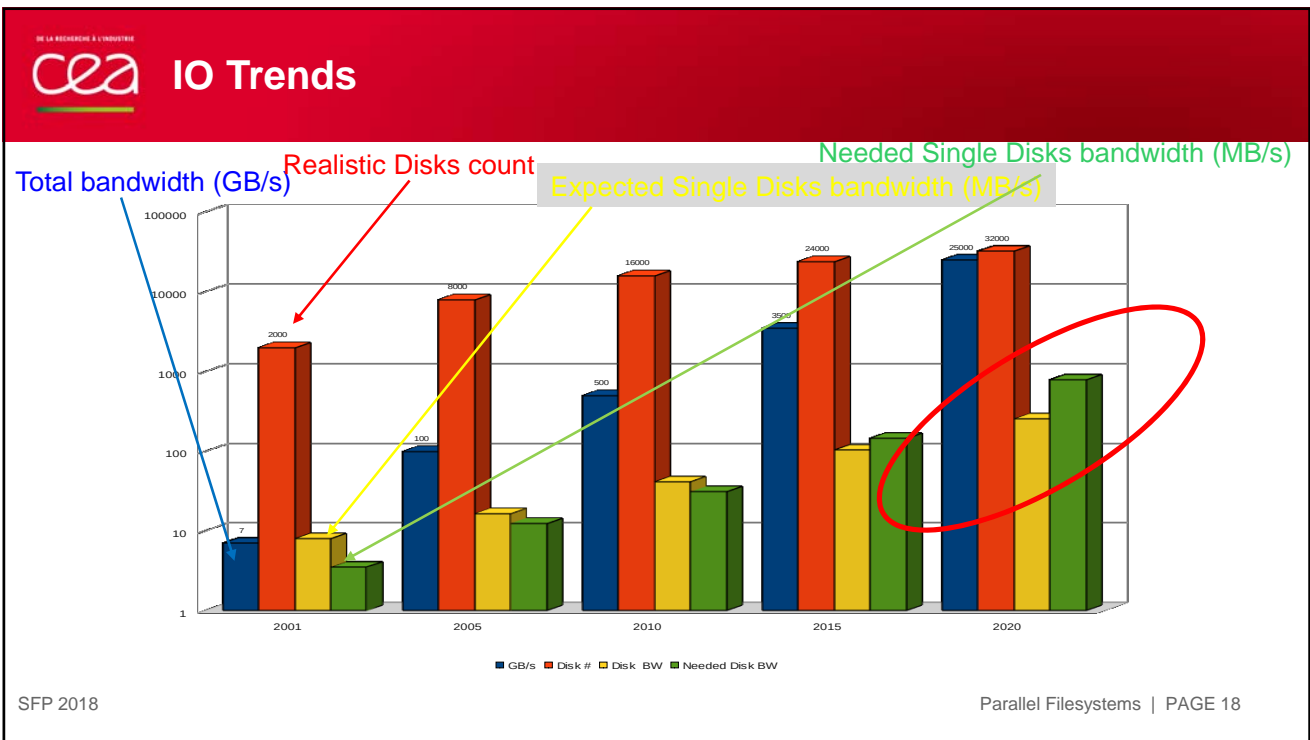
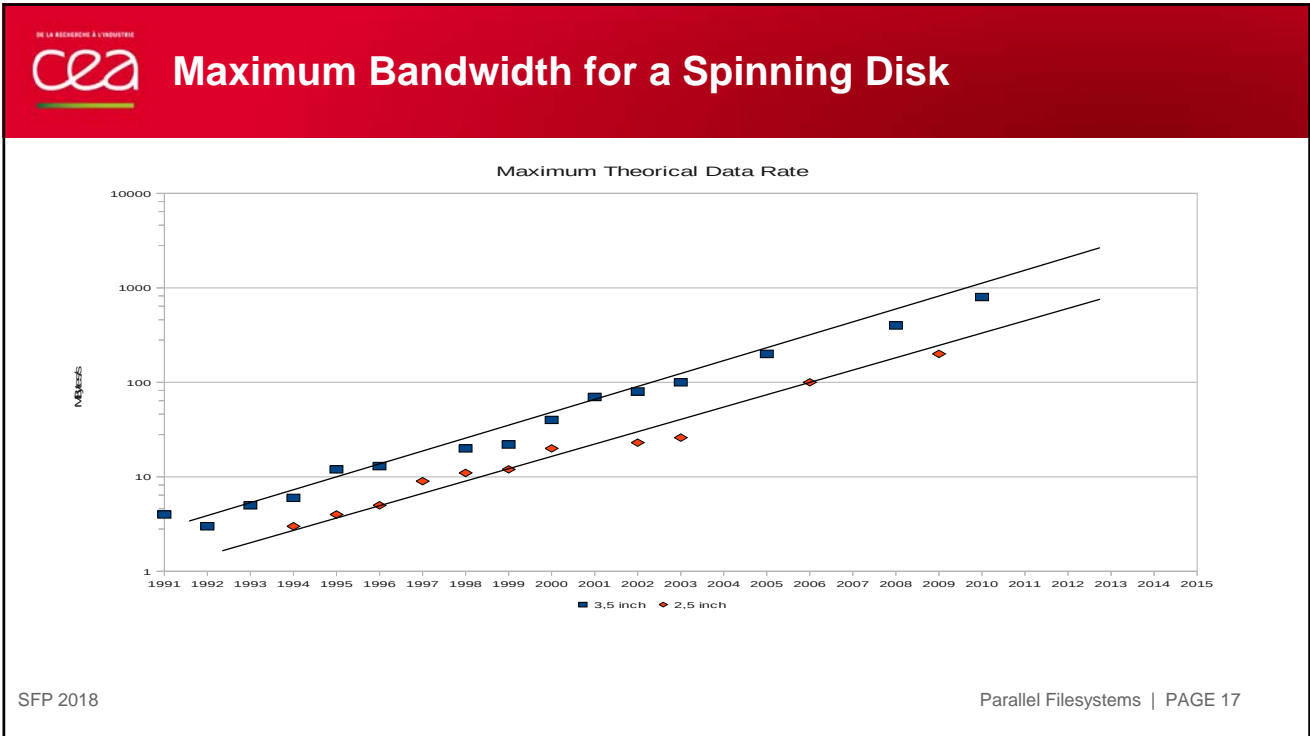
- ManyCore architecture (ratio GB/Thread decreases)
- Memory starvation for operating system

SFP 2018 Parallel Filesystems | PAGE 15

**How to get filesystem performances?**

| PAGE 16







2018+

## Spinning storage bandwidth is not enough

- Massive use of Flash storage (1 TB/s in 2019)

## Embedded model (Active Storage)

- IO servers runs in RAID controllers
- Reduce IO servers costs
- Increase IO server efficiency
- Prototypes available in 2012

## Upcoming challenges for file systems

- Heterogeneity support
  - From flash to spinning disk
- High increase of parallelism: client # and threads/clients
- Meta-data scalability
  - User data-set size increases
  - Need for many more application meta-data associated to files

SFP 2018

Parallel Filesystems | PAGE 19

## Which Architecture for Storage in 2020?

| PAGE 20

DE LA RECHERCHE À L'INDUSTRIE  
**cea Servers**

### Today's architecture (block based) is too simple/low level

- Need for a new architecture with a larger global view
- Network object model
  - File server becomes an object server
  - Distributed network parity between objects servers

**Server**

**Application**  
File System DB

↓

POSIX  
File System  
Volume Manager  
Driver

↓ FC

**Storage Server**  
RAID  
Battery Backed RAM  
Cache

↓ SAS

**Devices**  
SAS Interface  
SMR, Mapping  
Cylinder, Head, Sector  
Drive HDA

➔

**Application Kinetic Library**

↑

Ethernet

↓

**Devices**  
Ethernet Interface  
Key Value Store  
Cylinder, Head, Sector  
Drive HDA

SFP 2018 Parallel Filesystems | PAGE 21

DE LA RECHERCHE À L'INDUSTRIE  
**cea Servers (Cont.)**

### Meta-Data Scalability

- Meta Data need to be hosted by multiple servers
  - Need for distributed transactions
  - Need for distributed fault tolerance

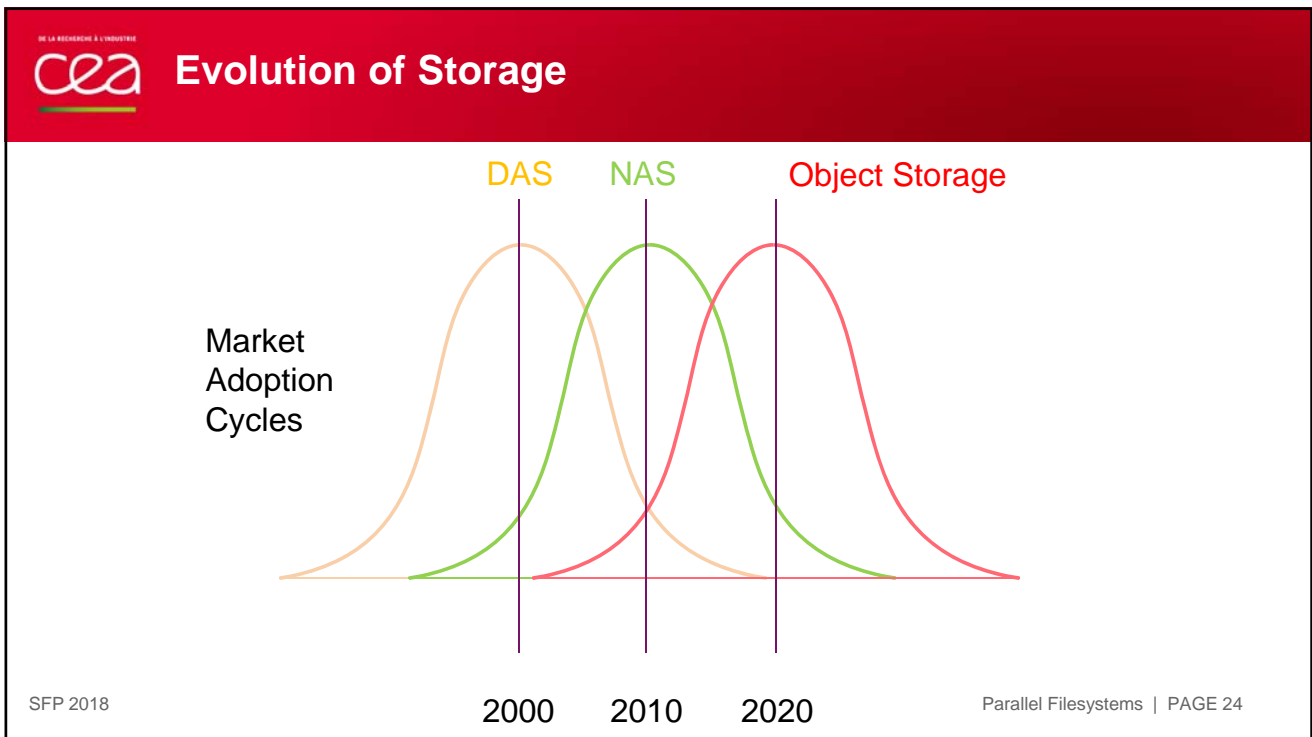
SFP 2018 Parallel Filesystems | PAGE 22

DE LA RECHERCHE À L'INDUSTRIE  
**cea** **Object Storage**

### New solutions arrives

- Seagate Kinetics
  - Ethernet connected disks
  - Dead ☹️
- OpenIO
  - Object Storage software
- New storage appliance
  - Standard disk
  - ARM based interposer
  - Open architecture

SFP 2018 Parallel Filesystems | PAGE 23



**cea Clients**

**Compute Node**

- Highly multi-threaded
- Few memory/thread

**File system client**

- Need memory for IO buffers

Need to introduce a mechanism to transfer IO to storage from compute nodes to an IO gateway

- Dynamic allocation of IO gateways
- Remote Direct Memory copy from compute node to limit memory use

2 tracks

- System: IO Proxy
- Applications : IO delegation

SFP 2018 Parallel Filesystems | PAGE 25

**cea Exascale Datapath**

**2018**

Compute Nodes  
FS Clients  
X 000

FS Servers

Storage Ctrl  
X00 Go/s

I/O delegation

**2020**

Compute Nodes  
X00 000

I/O Proxy  
FS Clients  
X 000

FS Servers  
Storage Ctrl  
X0 000 Go/s

SFP 2018 Parallel Filesystems | PAGE 26



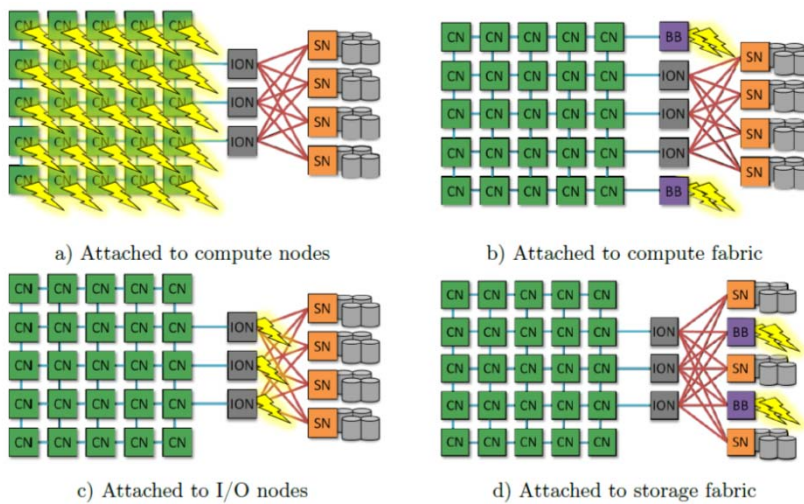
## Another Track: Burst Buffers

### Burst Buffer

- A burst buffer
  - Acts as a fast write-behind cache
  - Transparently migrates data from the burst buffer's fast storage to a traditional parallel file system
- Burst buffers rely on flash or NVM to support random I/O workloads that HDD-based file systems struggle with
- Specific API or FUSE for POSIX single node compliance
- Implementations
  - IME (DDN)
  - DataWarp (Cray)
  - FlashSystem (IBM)



## Different Burst Buffer Architectures



From Supercomputing Frontiers and Innovations



## Applications

### New IO paradigm

- Constraints from Posix interface need to be removed
  - No more possible to offer a free/fast global coherency to applications
- Applications and resource manager need to provide help to storage (hints)
  - Topology, access mode, ...
  - Real data use knowledge is within applications
- Working groups have started discussions on new IO API for applications (expansion phase)
  - EOFS EIOWG, Point2BDMC
- **Applications will have to change their IO interfaces to get all the performances**

SFP 2018

Parallel Filesystems | PAGE 29



## Summary

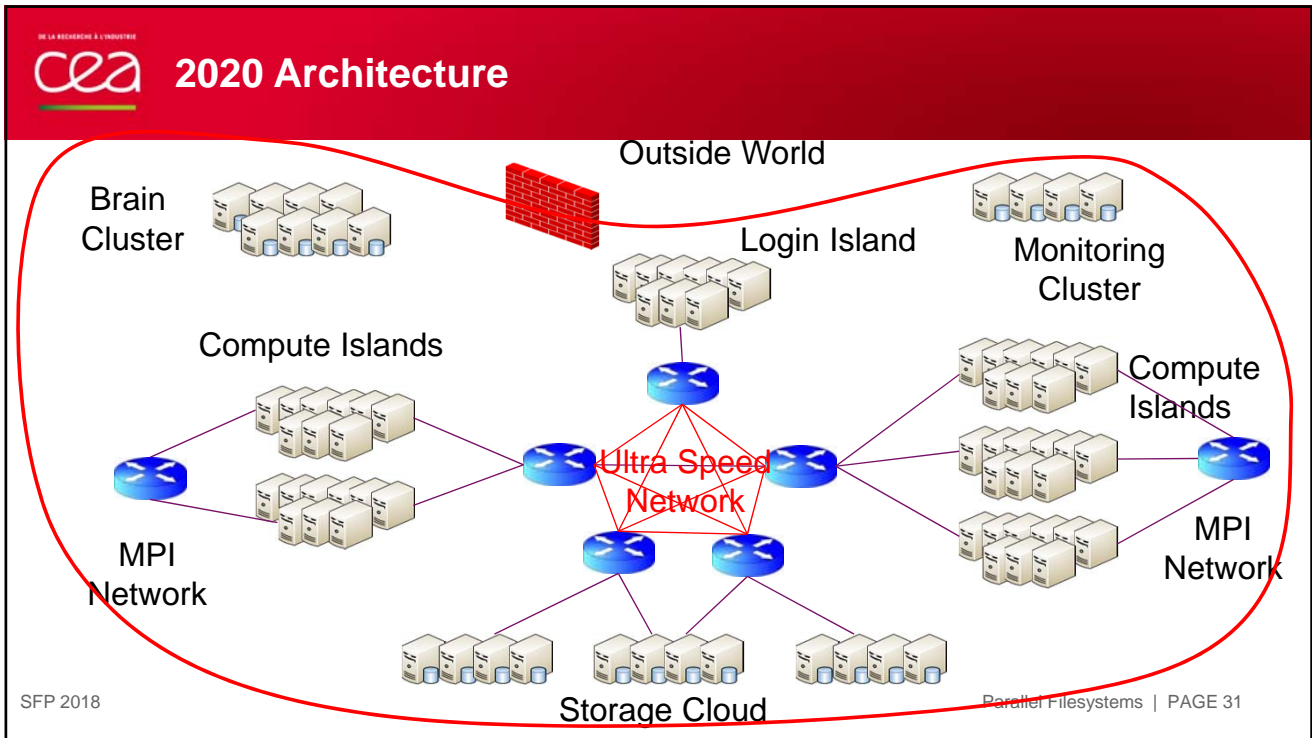
### IO challenges for future storage systems

- Data and Meta-Data management
- Hints from “those who know”
- New IO servers architecture
- New storage devices : Object Storage Devices
- I/O delegations

Multiple ways already exists  
Storage communities have started to work on solutions

SFP 2018

Parallel Filesystems | PAGE 30



**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