BLUE WATERS, A PETASCALE COMPUTER FACILITY, FOR THE RECONSTRUCTION OF CERN COMPASS-II DATA.

Marco Meyer

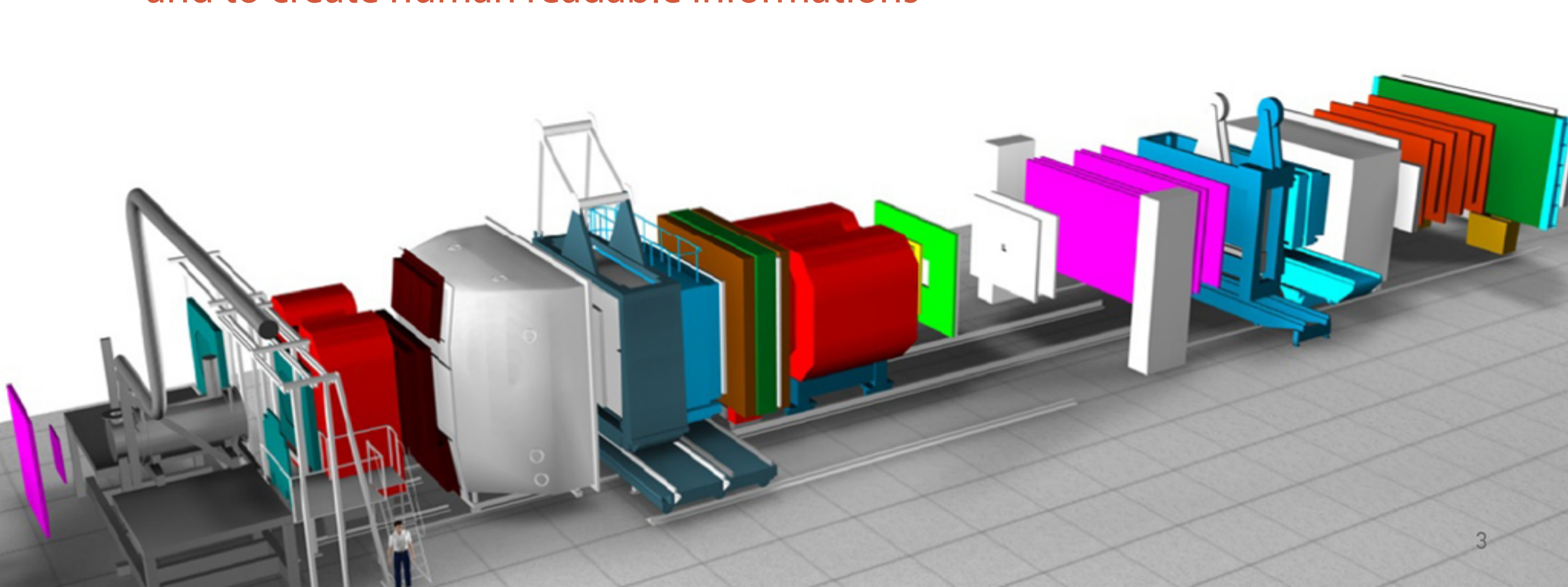
September 27, 2016 - 22nd International Spin Symposium

INTRODUCTION

- The COMPASS experiment at CERN
- The Blue Waters facility, on Campus at Urbana-Champaign.
- Roadmap and current status of performance evaluations
- Conclusion & Beyond the exploratory phase.

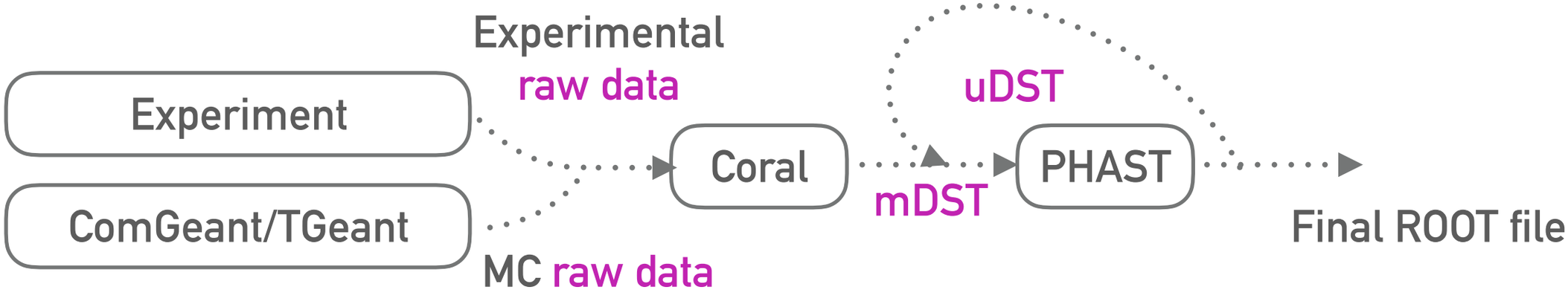
THE COMPASS EXPERIMENT AT CERN

- COMPASS is a high-energy physics experiment, on the Super Proton Synchrotron (SPS) at CERN (up to **190 GeV beam energy**)
- **Measurements performed** : Track reconstruction, energy measurement, particle identification
- **Complexe setup** which needs a lot of cpu time to decode, to reconstuction and to create human readable informations



THE COMPASS EXPERIMENT AT CERN

➤ COMPASS software chain

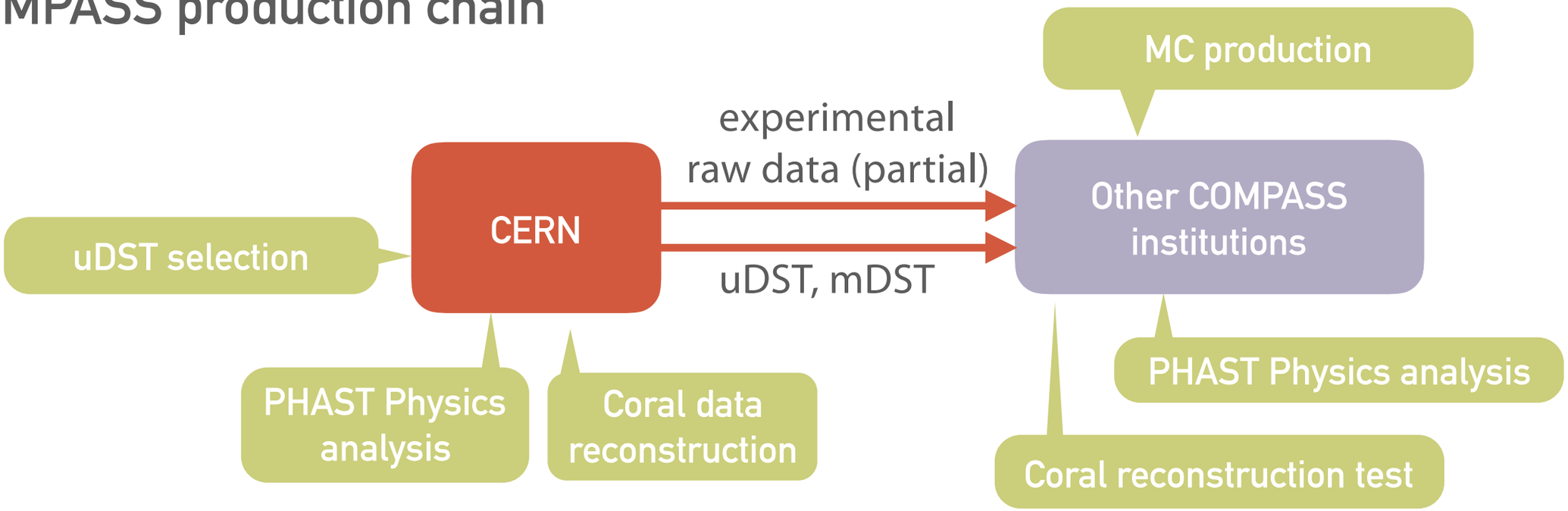


Software	Purpose	Input	Output
ComGeant	a MC spectrometer simulation package (Fortran)	MC generated events (Pythia)	MC Raw data
TGeant	Main MC spectrometer simulation package (ROOT/C++)	MC generated events (Pythia)	MC Raw data
Coral	Data reconstruction (ROOT/C++)	Raw data	**mini-DST***
PHAST	Physics analysis, detector efficiencies (ROOT/C++)	mini-DST/**micro-DST***	**ROOT file** (histograms, trees) or **micro-DST**

**DST = Data-Summary-Tree*
**micro-DST = subselection of miniDST*

THE COMPASS EXPERIMENT AT CERN

➤ COMPASS production chain



➤ Summary of the situation : **(2015 data ~ 744 TB of data)**

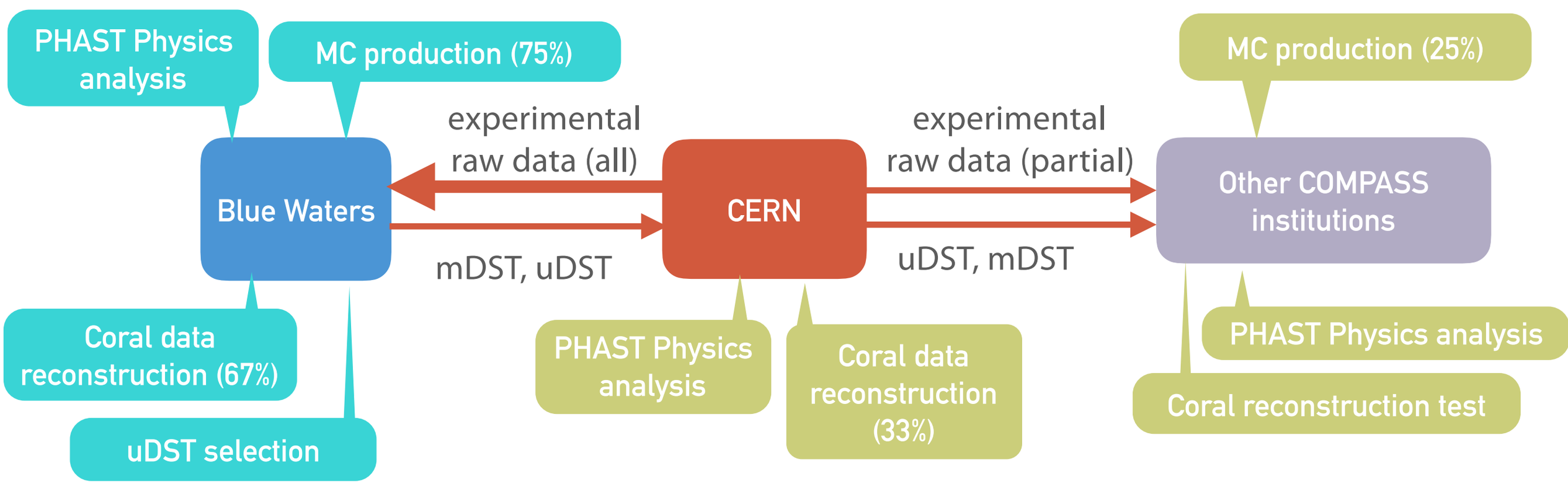
Element of the chain	Usual time unit (realtime processing)
Data taking	7 months
MC generation	4 x (7 months of data taking)
Coral reconstruction	2 weeks of data taking = 2 weeks of coral reproduction
uDST selection	2 weeks of data taking = 1/2 day for selection

THE COMPASS EXPERIMENT AT CERN

- **Blue Waters group project at COMPASS (for exploratory phase)**
  - Small working group started to explore Blue Waters for COMPASS

- **Problematic :** CERN Grid quota for COMPASS accounts is limited
  Main concerns are about data reconstruction (including upcoming 2016 data)

- **Proposed computing model with Blue Waters :**



Allocation granted: 40k node-hours
50TB on-line space
100TB near-line space

THE BLUE WATERS FACILITY

- **Blue Waters** : one of the world's most powerful computing center

- **Petascale computer**, located in Urbana-Champaign
  (PetaFLOPS = Peta-FLoating point Operations Per Second = measurement of the computer performances)

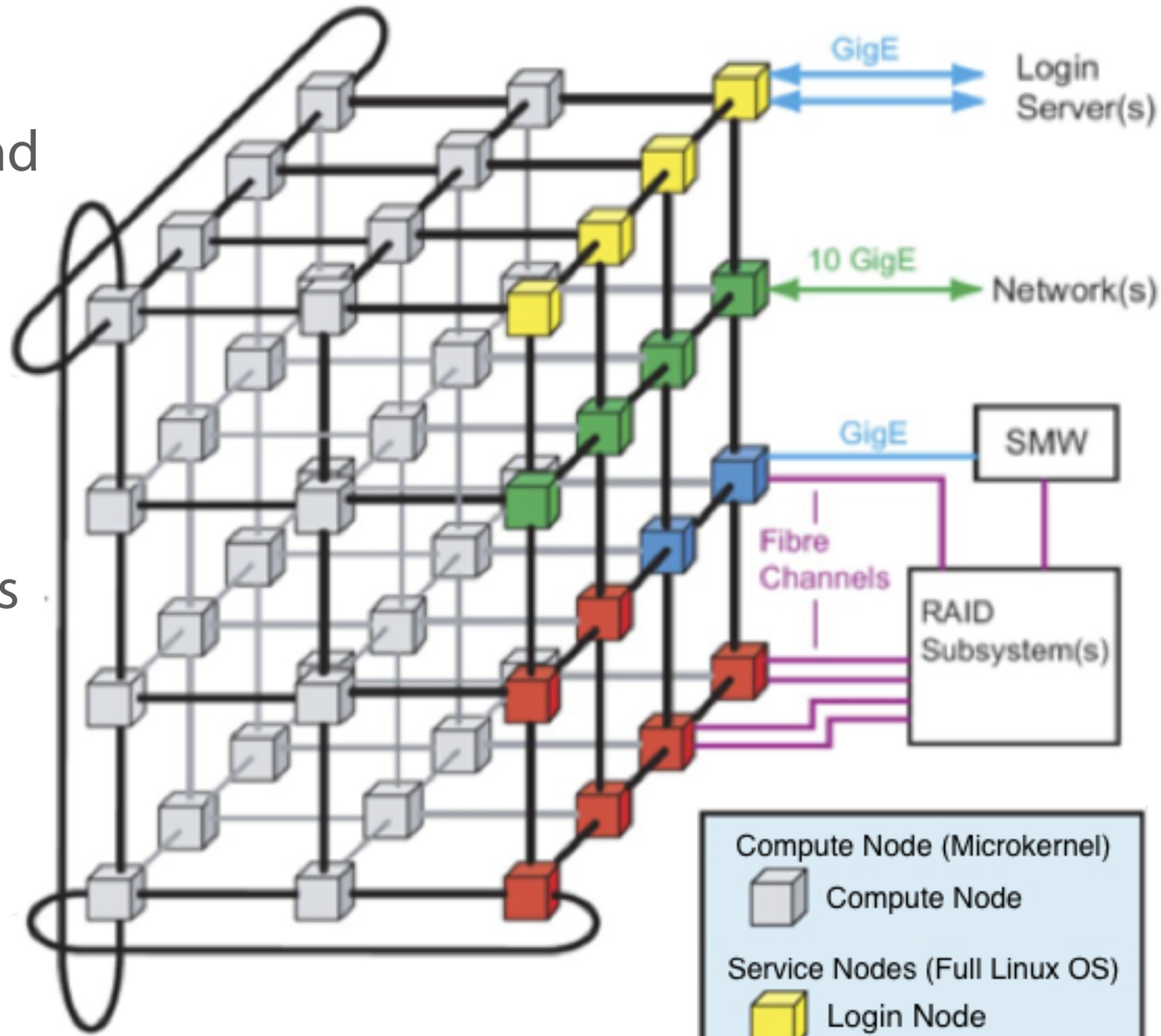
- A especially <u>well designed architecture</u>
  The keyword is the scalability : capacity to process data in parallel without any performance modification

THE BLUE WATERS FACILITY

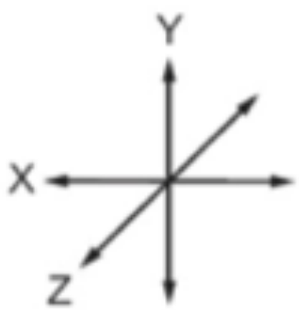
- **3D node architecture** (Cray Inc.)
  Elementary unit is **node-time** and not cpu-time

- **1 node** is a set of 16-32 CPUs

- Batch job grid = Compute nodes



Interconnection Network: 3D Torus in Each Dimension



THE ROADMAP AND CURRENT STATUS

- **May 2016 - Beginning of the exploratory BW project**
- Bandwidth measurement between CERN and BW
- *Group organization & COMPASS software*
- *First performance results on grid*
- *First big-data recontruction at BW*
- *November 2016 - End of the exploratory phase*

THE ROADMAP AND CURRENT STATUS

May 2016 - Beginning of the exploratory BW project

Bandwidth measurement between CERN and BW

Group organization & COMPASS software

First performance results on grid

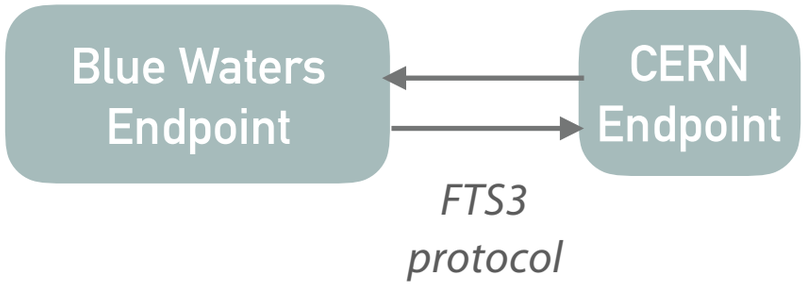
First big-data recontruction at BW

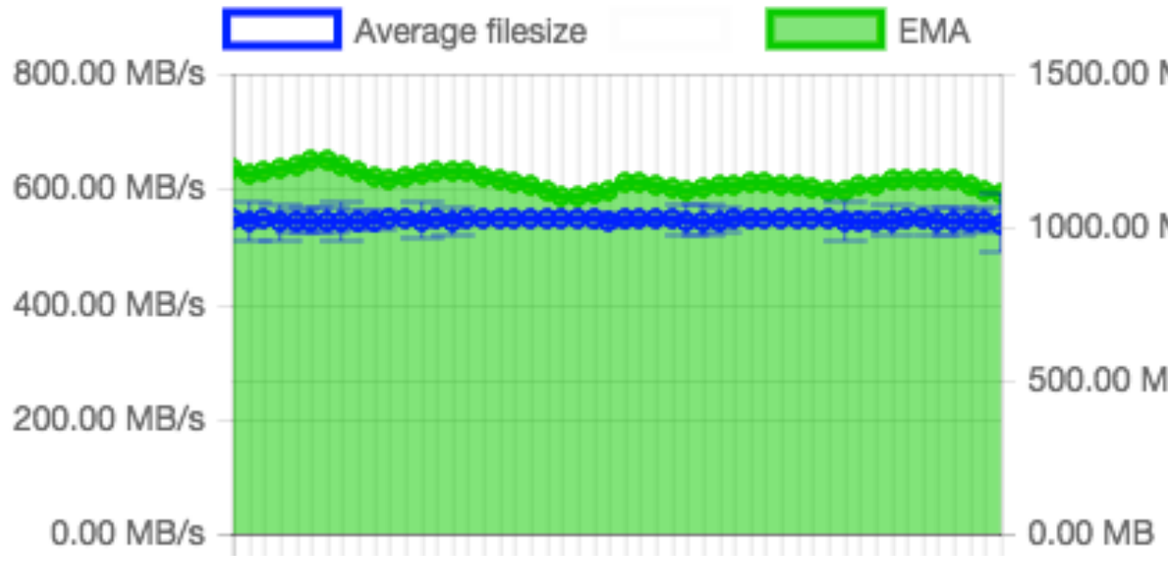
November 2016 - End of the exploratory phase

THE ROADMAP AND CURRENT STATUS

Bandwidth measurement between CERN and BW

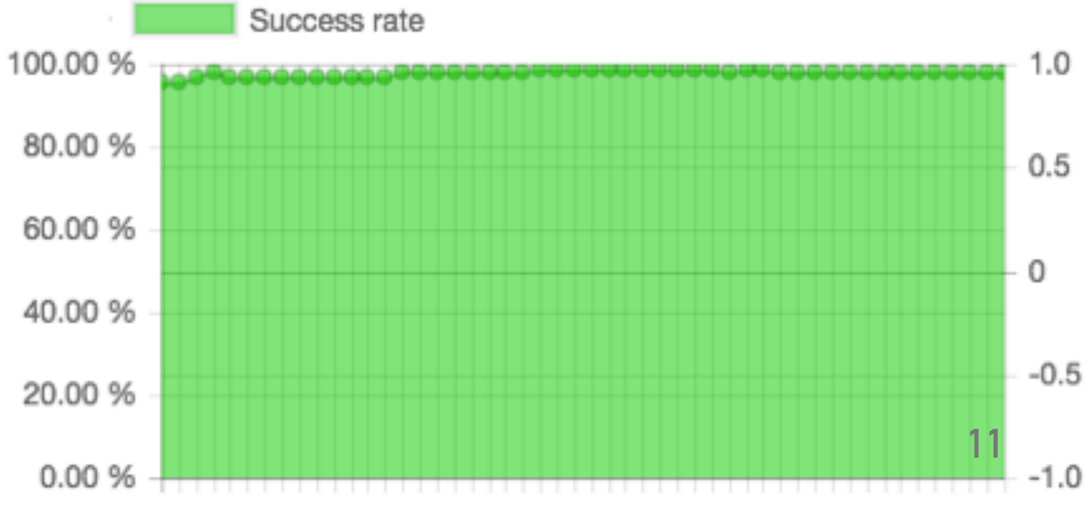
- Globus-Online tool : to transfer large amount of data between computing centers





- Bandwidth increases with number of files transferred at same time
  (**observed GigaByte rates : 800MB/s to 1200MB/s**)

- 2015 raw data ~ 744 TB
  transferred raw data ~ 160TB



THE ROADMAP AND CURRENT STATUS

May 2016 - Beginning of the exploratory BW project

Bandwidth measurement between CERN and BW

Group organization & COMPASS software

First performance results on grid

First big-data recontruction at BW

November 2016 - End of the exploratory phase

THE ROADMAP AND CURRENT STATUS

Group project common directory :

- Include **COMPASS files** needed for data reconstruction
- Store **pre-compiled COMPASS software**
- Group environment configuration

Goal of this organization :
Be well organized to be prepared for the next round (full BW proposal)

COMPASS software

- **We installed on Blue Waters all required :**
  - Physics libraries needed (LHAPDF, CLHEP, ..)
  - Generic particle-physics software (ROOT, Geant3, Geant4,..)

THE ROADMAP AND CURRENT STATUS

COMPASS software

Software	Purpose	Installation	First use	Tested on grid	Scalability test
ComGeant	Previous MC generator (Fortran)	✅	✅	Not done yet	Not done yet
TGeant	Current main MC generator (ROOT/C++)	✅	✅	✅	✅
Coral	Data reconstruction (ROOT/C++)	✅	✅	In progress	In progress
PHAST	Physics analysis, detector efficiencies (ROOT/C++)	✅	✅	✅	✅

➤ **Important information** : None of our software are multi-threaded

THE ROADMAP AND CURRENT STATUS

May 2016 - Beginning of the exploratory BW project

Bandwidth measurement between CERN and BW

Group organization & COMPASS software

First performance results on grid

First big-data recontruction at BW

November 2016 - End of the exploratory phase

THE ROADMAP AND CURRENT STATUS

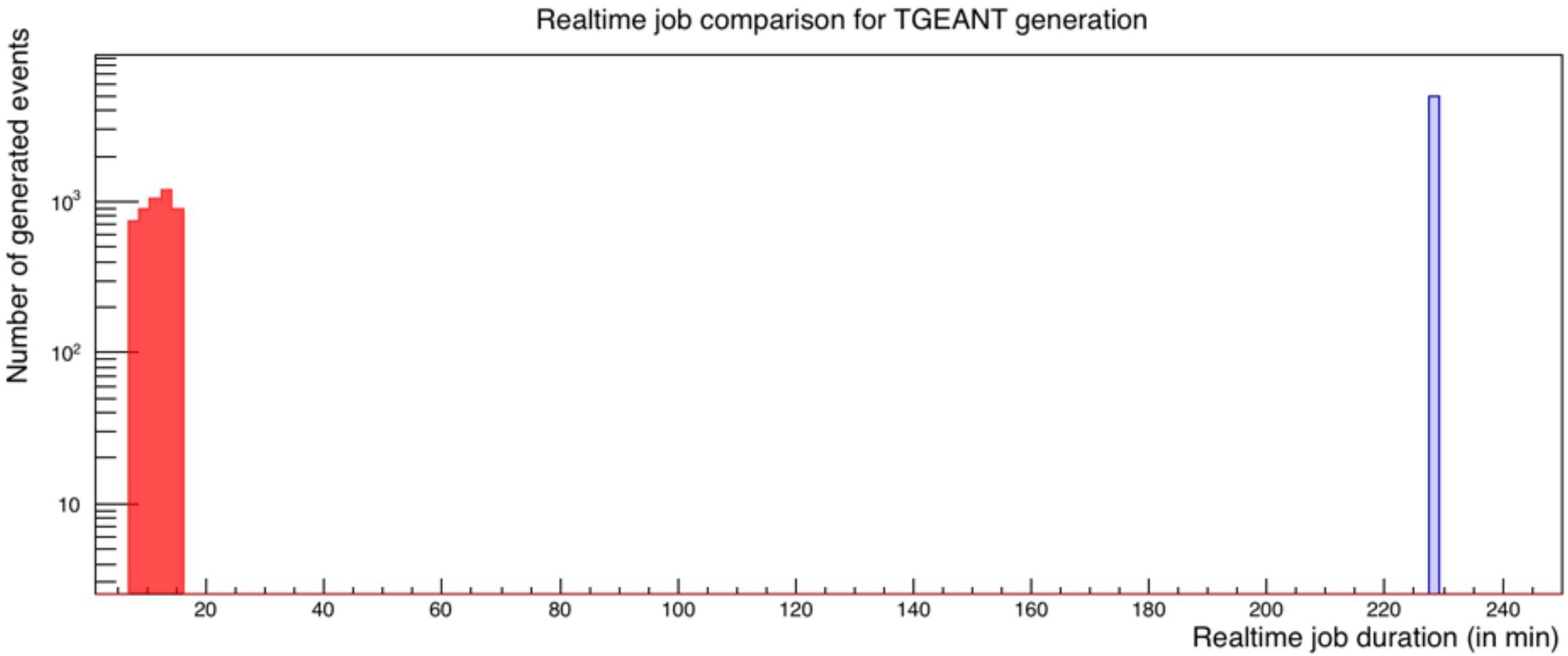
First results with PHAST on grid (based on uDST production)

- Based on an set of 4249 input mDST
- At CERN ~ 10h ; At BW ~ 8 min (Thanks to the scalability)
  Improvement factor ~75; for this specific set

THE ROADMAP AND CURRENT STATUS

First results with TGeant on grid (based on the 2015 setup of COMPASS)

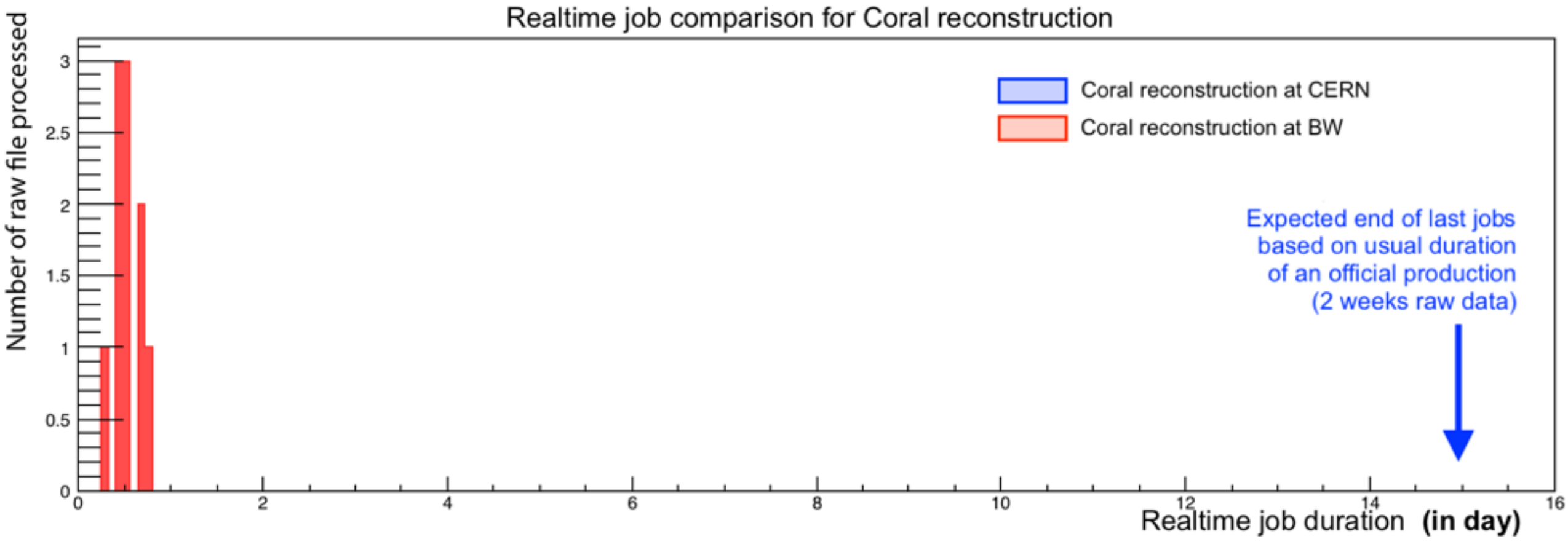
- Based on a set of 5000 generated events
- At CERN ~ 3-4h ; At BW ~ 20 min (Thanks to the scalability)
  Improvement factor ~12; for this specific set



THE ROADMAP AND CURRENT STATUS

First results with CORAL on grid (based on the COMPASS production model)

- Only few raw data files processed at BW
  (Scalability test to be performed)



THE ROADMAP AND CURRENT STATUS

May 2016 - Beginning of the exploratory BW project

Bandwidth measurement between CERN and BW

Group organization & COMPASS software

First performance results on grid

First big-data recontruction at BW

November 2016 - End of the exploratory phase

THE ROADMAP AND CURRENT STATUS

First big data production

- **Final test**: to use left over node-hours (~ 30k node-hours)

- Production, based on the option files used at CERN for the official production

- **Set of data to use** : 90k raw files to process (~ 95TB)

- **Main Goal** : Process data in 10-20 hours, instead of two weeks at CERN
  - Check the matching with the COMPASS production

THE ROADMAP AND CURRENT STATUS

May 2016 - Beginning of the exploratory BW project

Bandwidth measurement between CERN and BW

Group organization & COMPASS software

First performance results on grid

First big-data recontruction at BW

November 2016 - End of the exploratory phase

CONCLUSION & BEYOND THE EXPLORATORY PHASE.

- The observed **data transfer rates** are **sufficiently high** to transfer an annual COMPASS data set in 12 days from CERN to Blue Waters.

- **COMPASS software for first tests are running fine.**
  Blue Waters is very promissing with impressive capabilities. Good candidate for data reconstruction. **Scalable architecture** is a great advantage.

- **Next step** : Mass production, to process ~2 weeks of COMPASS raw experimental data to claim a final word..

- **Outlook**: We have applied for more computing time on Blue Waters beyond November 2016

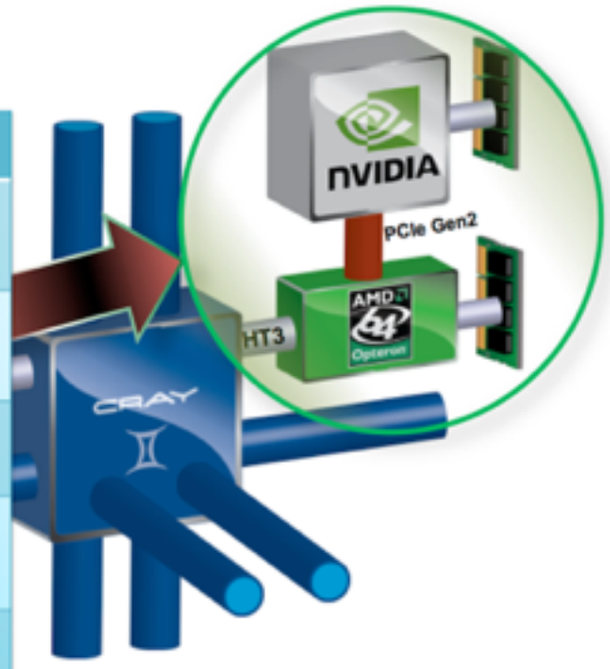
THANK YOU FOR YOUR ATTENTION

Many thanks to Blue Waters team and to all contributors

Marco Meyer - Blue Waters project for COMPASS-II data reconstruction
September 27, 2016 - 22nd International SPIN Symposium

THE BLUE WATERS FACILITY

XK7 Compute Node Characteristics	
Host Processor	AMD Series 6200 (Interlagos)
Host Processor Performance	156.8 Gflops
Kepler Peak (DP floating point)	1.32 Tflops
Host Memory	32GB 51 GB/sec
Kepler Memory	6GB GDDR5 capacity > 180 GB/sec
Number of Cores	32

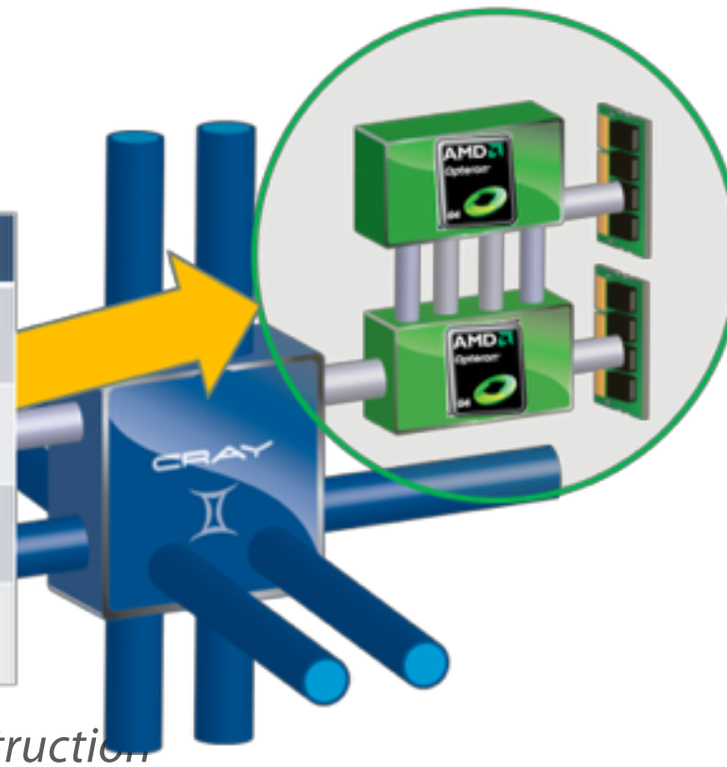


- **4,228 Cray XK7 nodes**
  - 32 GB memory with
  - NVIDIA graphics processor acceleration.

- **22,640 Cray XE6 nodes**
  - with each 64 GB memory

XE6 Compute Node
- Dual-socket AMD-Opteron
- 4x channel 1600 DDR3 memory
- High speed HT3 network link
- Upgradeable
- Blend with XK6 GPU systems

Node Characteristics	
Number of Cores	16
Peak Performance	313.6 Gflops/sec
Memory Sizes Available	64 GB per node
Memory Bandwidth (Peak)	102.4 GB/sec



THE BLUE WATERS FACILITY

Near-line Storage

Archive Software	HPSS
Online disk cache	1.2 PB
Aggregate Bandwidth to tape	58 GB/s
5 year capacity	380 PB

Interconnect

Architecture	3D Torus
Topology	24x24x24
Compute nodes per Gemini	2
Peak Node Injection Bandwidth	9.6 GB/s

Online Storage

Total Usable Storage	26.4 PB
Total Raw Storage	34.0 PB
Aggregate Measured I/O Bandwidth	> 1.1 TB/s

File System	Size (PB)	# of OSTs
home	2.2	144
projects	2.2	144
scratch	22	1440

XE Compute Node

AMD 6276 Interlagos Processors	2
Bulldozer Cores*	16
Integer Scheduling Units**	32
Memory / Bulldozer Core	4 GB
Total Node Memory	64 GB
Peak Performance	313.6 GF
Memory Bandwidth	102.4 GB/s

XK Compute Node

AMD 6276 Interlagos Processors	1
Bulldozer Cores*	8
Integer Scheduling Units**	16
Memory / Bulldozer Core	4 GB
Node System Memory	32 GB
GPU Memory	6 GB
Peak CPU Performance	156.8 GF
CPU Memory Bandwidth	51.2 GB/s
CUDA cores	2688
Peak GPU Performance (DP)	1.31 TF
GPU Memory Bandwidth (ECC off)***	250 GB/s

MULTI-THREADING

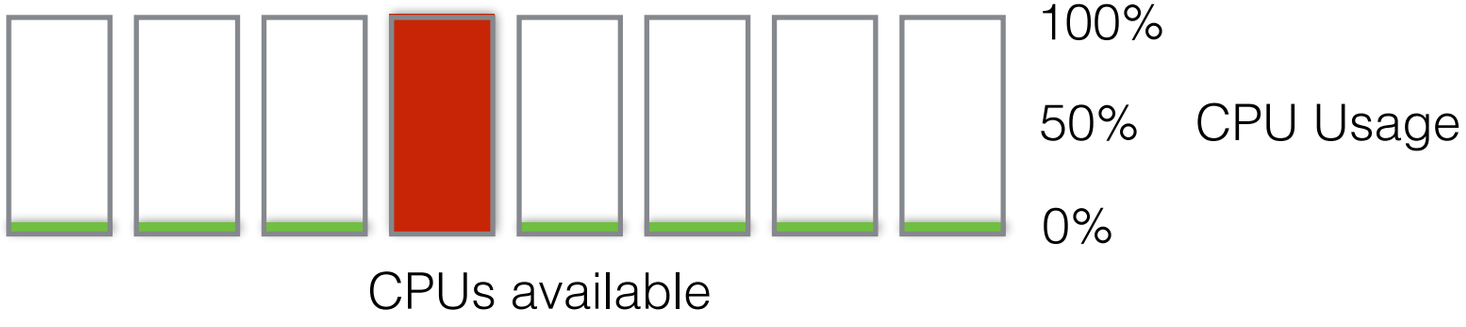
➤ No multi-threading means : **N files to process = only 1 cpu used**

➤ Multi-threaded software means : **N files are spread over N cpus**

1 file = a set of physics events

REALTIME VS CPUTIME

- e.g. 32 files; 32 cores (1 file = 1 cpu-hours)
- Situation A : (No multi-threading)



CPUtime = 32 cpu-hours

Realtime ~ CPUtime + Waiting-time

Realtime ~ CPUtime
(neglecting waiting-time)

- Situation B : (Including multi-threading)

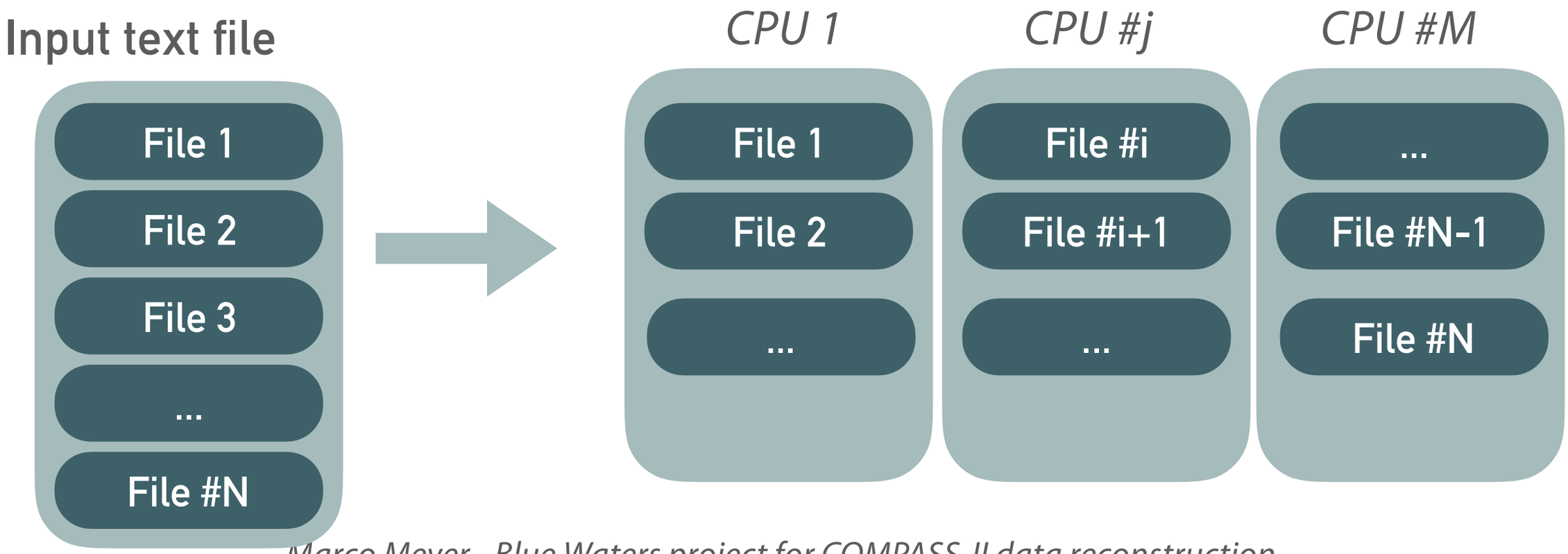
CPUtime = 32 cpu-hours

Realtime ~ CPUtime/32 + Waiting-time

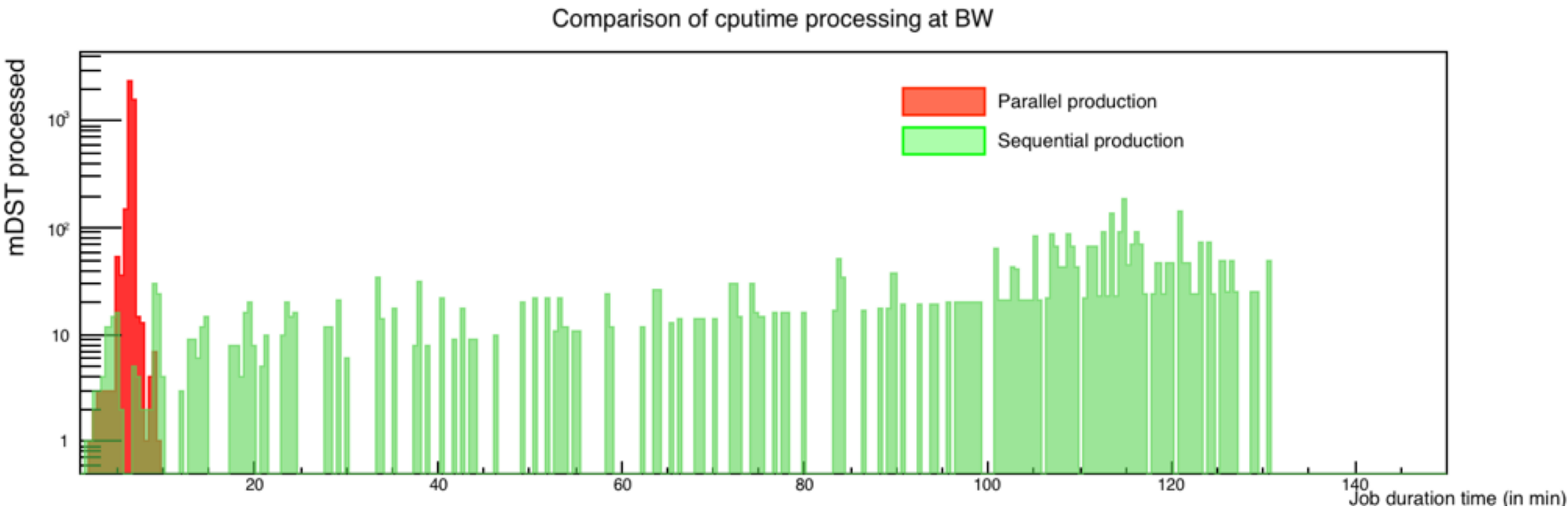
Realtime << CPUtime
(neglecting waiting-time)

PARALLEL COMMAND PROCESSOR (PCP)

- Script written in C++ (proposed by BW people)
  Based on OpenMPI (Multi-threading library)
- **Provide a text file as input** : one command per line
- PCP spreads commands over all the available CPUs by itself
  e.g. N files to process; and M cores available

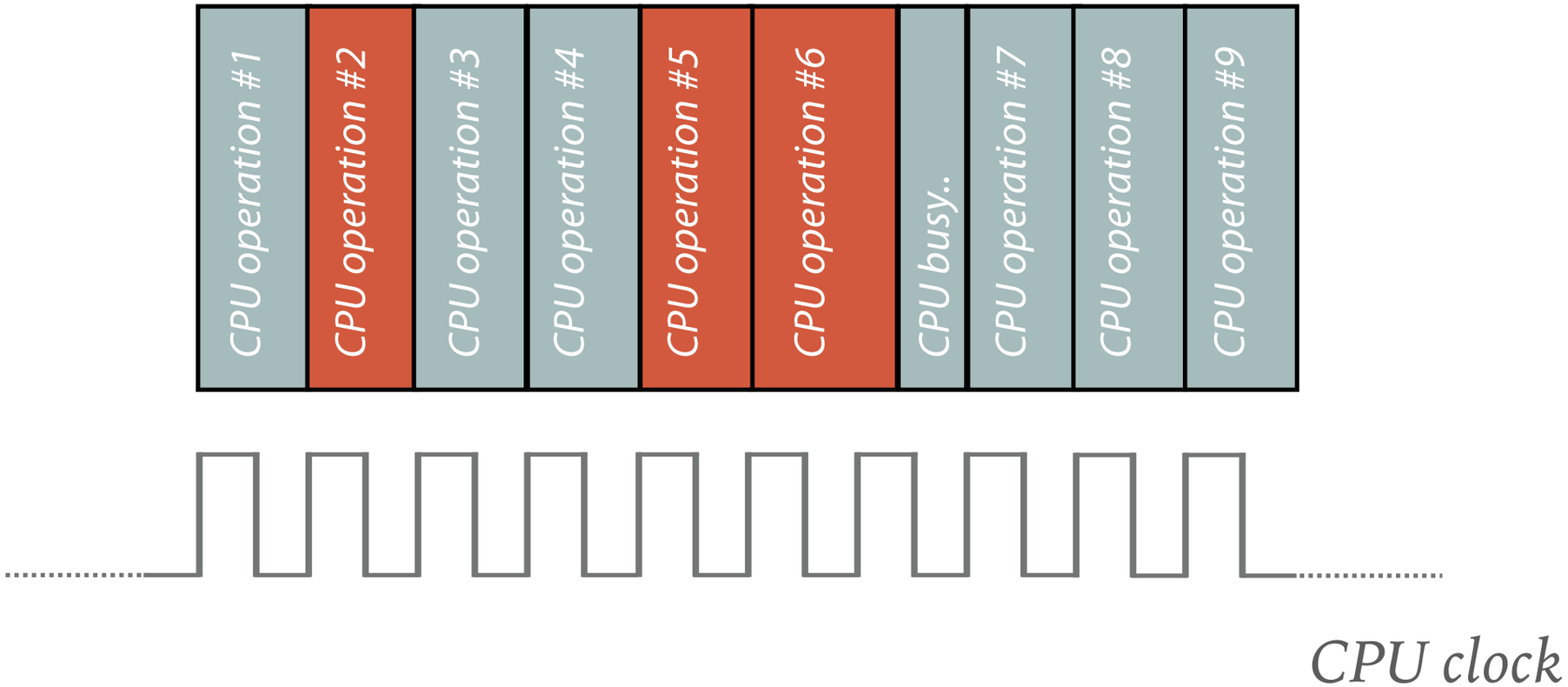


PARALLEL/SEQUENTIAL PROCESSING



CERN GRID MECHANISM

- Priority mechanism
- No dedicated CPU when running jobs..
- Fair-share CPU mechanism
- Limited number of job per collaboration



CERN PROXY RESTRICTION