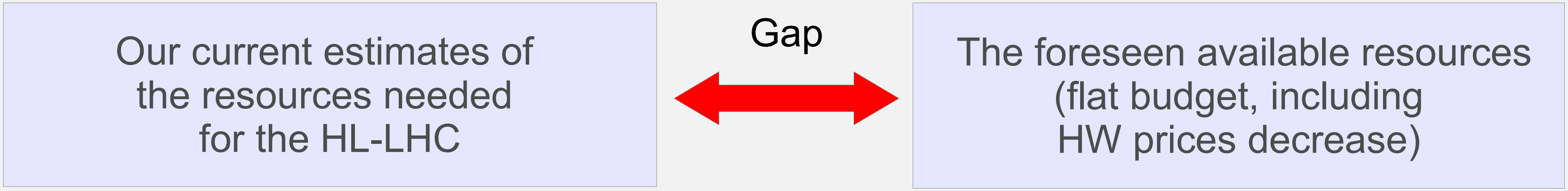
Towards designing a standardised cost model

WG “Systems Performance and Cost Modeling”

Catherine Biscarat (IN2P3, FR)
Jan Iven (CERN, CH)
Gareth D. Roy (University of Glasgow, UK)
Renaud Vernet (IN2P3, FR)

Joint WLCG/HSF workshop, Napoli, Italy, 2018, March 26-29

Motivation



Our current estimates of the resources needed for the HL-LHC

Gap

The foreseen available resources (flat budget, including HW prices decrease)

- We have to optimise the way we do the computing
  - This includes many elements (opportunistic resources, specialised HW, optimal SW, evolution of the workload…)
- We have to be able to quantify the impact of these changes on the infrastructure
  - And the cost need to be a design goal

- We have to establish a cost evaluation model
  - Effective whatever the model is / will be
  - We are not speaking only about pledges, but the overall cost of our infrastructure
  - It should take into account the main (sensible) components

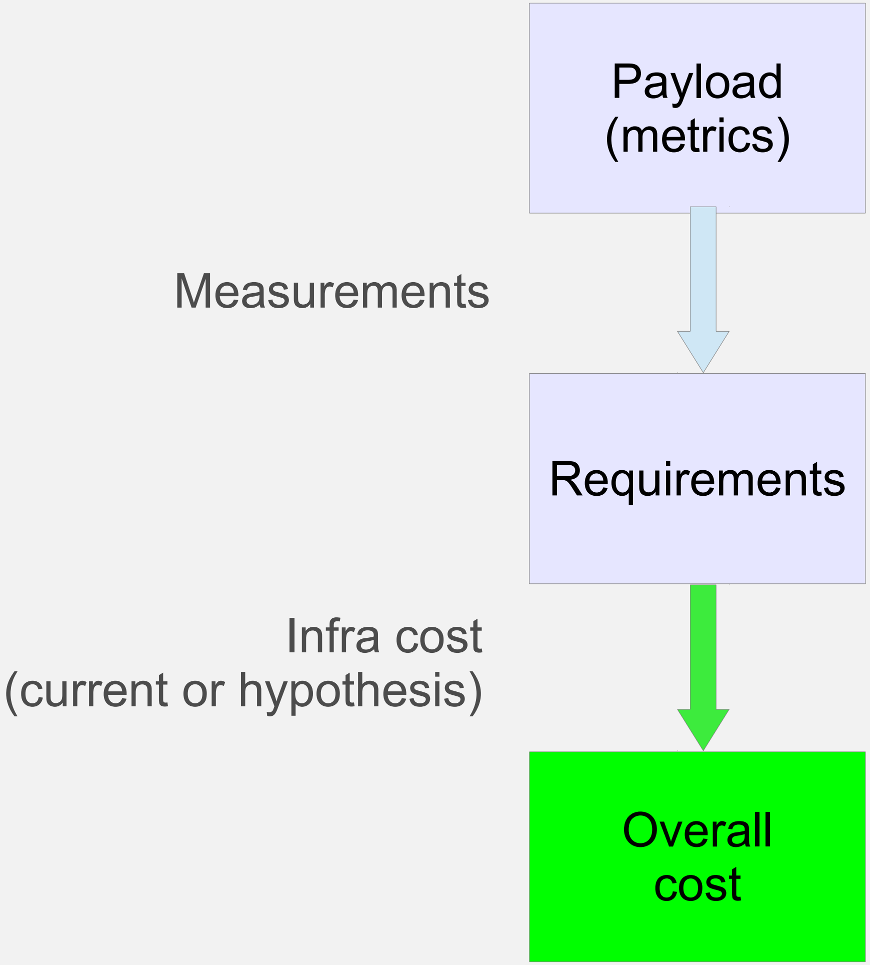
- As a start, we base this model on our knowledge of the current costs and computing model
- Sites and experiments could then easily estimate costs and do it in a coherent way.

Task – to draft cost evaluation process

Our task

- To outline a method to calculate the cost of hardware to expand the site capacity to hold the most relevant payloads
- This naturally breaks down into several components and should be modular.

Collaboration between infra. and experiments is crucial.



As a starting point

1st exercise

- Given the resource needs of a given workload, close to an existing one, a few sites describe how they would estimate the cost, based on their current infrastructure, to run this type of workload. Then the cost to run an additional 10^6 units of this workload within a year.

- The purpose of the exercise is
  - to understand what in the description of the workload is used ;
  - to identify what is missing ;
  - to discuss the approach to reach an estimate.

- Four sites volunteered. They are anonymised in the rest of this presentation.
- Only the 1 million job results are shown.

Payload description

Pseudo Pileup Digitization and reco job (experiment inspired)

Start with minimal a description → what additional information is missing ?
Assumption: very same behaviour across different architectures

Data in	1.5 GB individual data for a job & access to 30 GB shared by many jobs	Storage (incl. scratch)
Data out	6 GB (ESD) + 700 MB (AOD) with a data retention time of 12 months	
Intermediate data that during the job is written and read back	12 GB	
Total block reads/writes	78 GB in and 19 GB out	Disk speed
Processing needs	100k sec CPU time job on 8 cores of 10 HS06 each	Compute
Memory needs	16 GB	

Considering recent HW (a current workload)

Idea: with our knowledge of HW prices, we checked if it can handle this payload (CPU)

Total storage per WN (40 HT-core, e.g. 5 jobs) = 131 GB
already in the WN specification

Already in WN spec': 3-4 GB / HT-core

Current payload
& current HW
It fits

Data in	1.5 GB individual data for a job & access to 30 GB shared by many jobs
Data out	6 GB (ESD) + 700 MB (AOD) with a data retention time of 12 months
Intermediate data that during the job is written and read back	12 GB
Total block reads/writes	78 GB in and 19 GB out
Processing needs	100k sec CPU time job on 8 cores of 10 HS06 each
Memory needs	16 GB

Storage
(incl. scratch)

Disk speed

Compute

First order approach

Storage (TB) needed for data retention

CPU (HS06) needed for processing

Data in	1.5 GB individual data for a job & access to 30 GB shared by many jobs	Storage (incl. scratch)
Data out	6 GB (ESD) + 700 MB (AOD) with a data retention time of 12 months	
Intermediate data that during the job is written and read back	12 GB	
Total block reads/writes	78 GB in and 19 GB out	Disk speed
Processing needs	100k sec CPU time job on 8 cores of 10 HS06 each	Compute
Memory needs	16 GB	

Resulting costs

STORAGE

Payload inputs

- Retention data length
- Capacity to store per job
- Number of jobs

} 6.7 PB to store for 1 year

Site input

- Price per TB/year

Cost (1M jobs, 1 year)

Site	Storage
A	134 k€
B	250 k€
C	137 k€
D	220 k€

COMPUTE

Payload inputs

- Ncores
- HS06 / core
- Length of a job

} 254 kHS06.year to provide

Site input

- Price per HS06/year

Cost (1M jobs, 1 year)

Site	CPU
A	405 k€
B	810 k€
C	514 k€

Price per unit at sites

It is not enough to buy boxes
Prices / unit (TB or HS06) depend strongly on the components included in the calculation

- HW length and renewal (prices integrated over installed HW, or only recent HW)
- Rack
- Power outlets (PDU)
- Network hardware (switch+router port)

- Electricity
- PUE

- Facility/building
- Manpower (hardware support, config mgmt, storage management, network team)

Local configuration with same HW
- Storage: overhead → traduce in quality of service
- CPU: SL6/CC7, exec. points

Integration cost

- Three sites include integration costs.
- One of them gave the breakdown.
- Those ingredients/numbers depend on the DC configuration

- In this example:
  - Cost of the installation by the vendor
  - Cost of a water cooled rack
  - Cost of the network
    - No upgrade of the current network
    - Elements to reach the main rooter

With recent HW (dense for disk)

HW	Installation	To be added to bare metal cost
CPU	Each WN: 1 Gbps	+ 18.0% (rack > network > instal.)
Disk	Each unit (160 TB): 10 Gbps	+17.7% (rack > network > instal.)

- Equipment cost decreases with density (with equivalent network capacity)

Power consumption cost

And one has to power up the servers

- One (/4) sites pay for the electricity
  - Someone still pays for it, for doing Sciences

- Two approaches (very similar HW)

  1 - Consider only recent HW
  - PUE not included (1.1)

  2 - Consider the HW already installed
  - PUE is included (1.5-1.8)

  - In these examples: price/kWh is “low”

HW	1 – Recent HW (no PUE)	2 – Integrated HW (with PUE)
HW power consumption		
CPU	0.4 W/HS06	1.3 W/HS06
Disk	3.0 W/TB	12.0 W/TB
Cost to be added to integrated cost		
CPU	+23%	+33%
Disk	+14%	+27%

- Disk power consumption decreased over years (density effect).
- CPU consumption is more stable

Adapting HW (proc. memory)

- One site did the exercise to adapt its current CPU servers to the payload
  - Memory per HT-core

CPU	4 GB memory	2 GB memory
price	514 k€	-18%

Not included in this exercise

Stored elsewhere (e.g. another payload output) → we have to build a full picture (mixing payloads)
30 GB shared space already available

Network: we have to calculate the network capacity needed
→ storage boxes / core network / CPU boxes
We simply did check there was no bottleneck in existing infra.

Data in	1.5 GB individual data for a job & access to 30 GB shared by many jobs	Storage (incl. scratch)
Data out	6 GB (ESD) + 700 MB (AOD) with a data retention time of 12 months	
Intermediate data that during the job is written and read back	12 GB	
Total block reads/writes	78 GB in and 19 GB out	Disk speed
Processing needs	8 cores with 100k sec CPU time and 10 HS06 each	Compute
Memory needs	16 GB	

Results of first exercise

This very 1st exercise

- Four sites have evaluated the cost of the given payload
  - They have shared details on costs regarding bare metal, integration, power consumption
- The process in the cost evaluation is different from one site to the other
- Some characterisations of the payload were used to check if the costed infra. is capable to handle such payloads (already covered by current spec.)

Discussed as important to evaluate a global cost

- Job efficiency (negligible here), job scheduling efficiency, licence cost per job
- Network (file access), IO
- Mix of payload
- And many more when it will come to tapes (tape, robot, library, licence, buffer) – not so cheap
- Manpower

To go further

Remember, the final goal is

- not to compare the cost of different sites (we did follow up on the few differences that we have. They proved to be powerful indicators for differences in QoS, integration cost...)
- to be able to understand the implication (in term of cost) of the computing model changes we envision for the HL-LHC
  - Ideas are infinite, the budget is not

To do so, we have to

- Build a tool calculating the cost of any payload based on a job description
- Collaborate closely between experiments and resources providers (HW price)

Today

- We presented you our first findings playing with a given payload
- We need your inputs on:
  - The relevant ingredients to be included
  - The level of details to include, how much shall we expand the model ?
  - The value of such a tool for you (experiments ? sites ?)

Please participate to this afternoon discussion session !