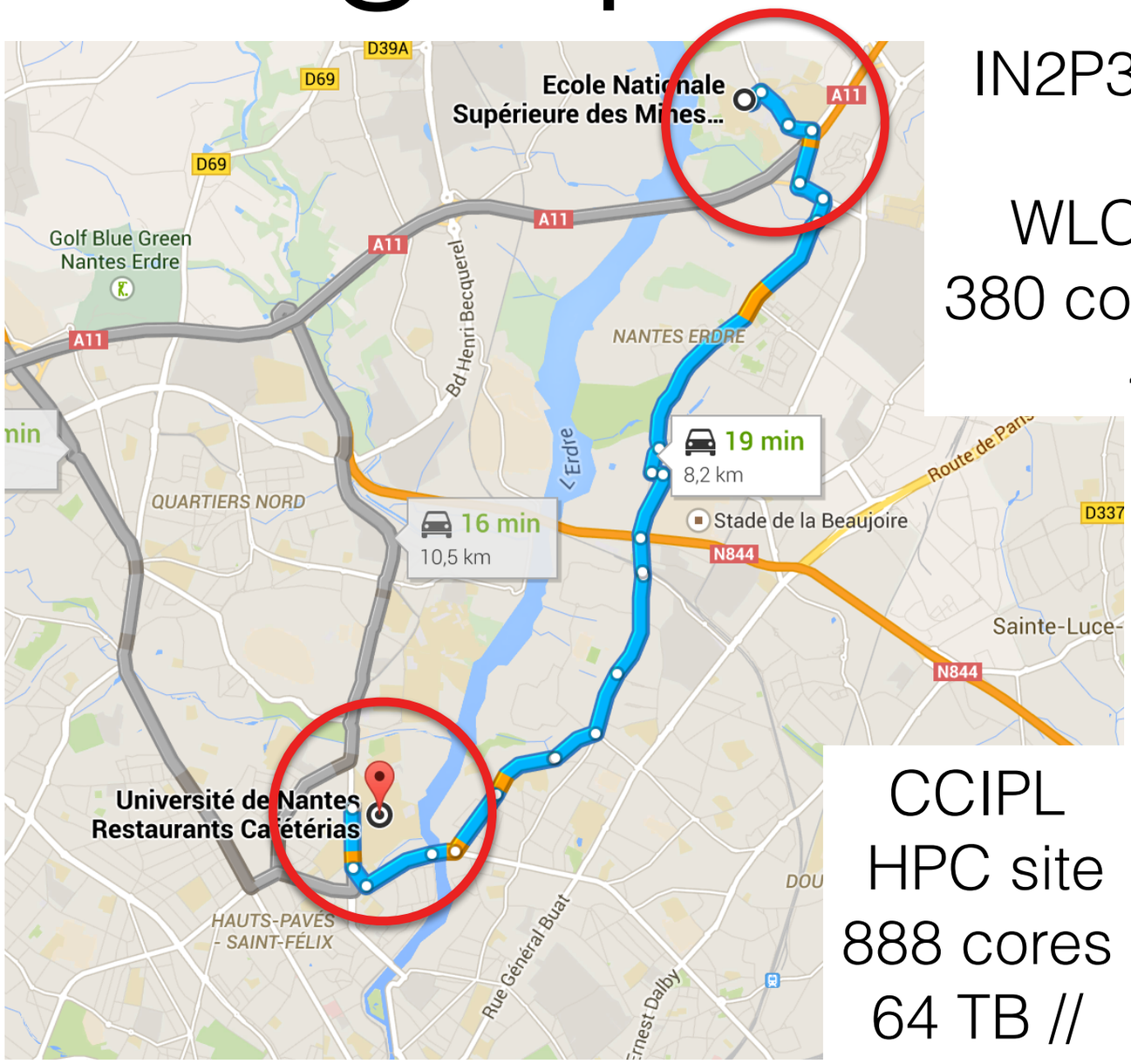
Collaboration project with HPC cluster in Nantes

L. Aphecetche

Geographical context



IN2P3-SUBATECH
T2
WLCG Grid site
380 cores (5k HS06)
434 TB

CCIPL
HPC site
888 cores
64 TB //

- Same city, two small data centers serving different communities (HEP=grid; others=HPC)

Financial context

- Easier to get money to start projects than to sustain them...
    - As a consequence Subatech T2 has had a ~flat ressource profile since 2012
- End of 2012, approached a local HPC cluster team to propose a joint fund hunting, aka joint project
    - turned out they, as well, were convinced that their financial future was in collaborating with others
- Around the same time a new data center (university wide) was decided
    - is now built and being setup
    - will host the new HPC cluster

CCIPL

- Centre de Calcul Intensif des Pays de la Loire (Intensive Computing Center of "Pays de la Loire")
  - meant to be a research (i.e. academic only) tool for the whole region
  - serves different communities in need for parallel computing (earth/planet science, chemistry, ...)
  - Currently 888 cores (9.4 Tflops), IB network, 64 TB parallel filesystem

Joint venture

- The idea was to make a fund request (as part of a bigger request, CPER (*)) in order to
  - renew completely the CCIPL
  - secure some funds for the Alice T2 and allow for some ressource increase (both « static » and opportunistic ones)
  - mutualize ressources as much as reasonably achievable (if not mutualized we cannot get opportunistic ressources)
- Overall budget 1M€ (700 k€ CCIPL 300 k€ T2), for the period 2015-2020 (i.e. Run 2). Granted end of 2014. Now we have to make it work ;-)
- Given the differences of need/human ressources/wishes/constrains, we settled on :
  - storage (~2/3 of our foreseen budget) to remain hosted at Subatech
    - disk procurements are fully under our control
  - CPU to be hosted at the new University Data Center
    - T2 CPU procurements will be synchronized with those of CCIPL

(*) Contrat de Plan Etat-Region

Spending profile (very preliminary (*))

[k€ TTC]	2015	2016	2017	2018	2019	2020
CCIPL		450		250		
T2 CPU		50		50	50	
T2 disk	90		50			
T2 network	10					

2016 procurement will be launched end of this year so the installation is foreseen Q1 2016

(*) we don't know yet if it fits the foreseen funding profile…

Estimated capacity increases (*)

	2015	2016	2017	2018	2019	Final capacity
T2 CPU		+3200 HS06		+4000 HS06	+4400 HS06	16 kHS06
T2 disk		+360 TB		+280 TB		
rel. to current		x1.5		x2.5	x3	1 PB

(*) caveat : using here the LCG-FR costs for those estimates.
CPU machines for CCPIL might be a bit less cost effective.
On the contrary disk estimates might be a bit on the pessimistic side

Towards a plan

- the new data center is being setup and a lot is still to be defined by its « clients »
- CCIPL (and T2) is a big client, so it has a major voice
  - In addition the (yet to be hired…) head of the DC will also be the head of CCIPL ;-)
- We discussed a bit with CCIPL so far, trying to identify hard points (*)
- **Need Alice input and expertise to be sure we're heading in the right direction**. Hence this talk…

(*) see also Jean-Michel presentation at last T1/T2 workshop at Tsukuba
http://indico.cern.ch/event/274974/contribution/95/material/slides/0.pdf

Clusters

- CCIPL currently envisions 3 clusters
  - 1 IB cluster for MPI jobs : not our business
  - 1 « regular » (ethernet) cluster for SMP jobs
  - our T2 cluster (ideally of the very same kind as the SMP one)
    - our needs are ~ the same here
    - machines with ~8 cores (they don't necessarily want the highest core count machines)
    - they need more memory per core (4GB) than we officially do (2GB), but that's OK I guess ;-)
    - OS : Linux with long term support (so RHEL or SUSE basically)

Does Alice have more requirements for the worker node machines we should take into account?

Feeding the system with grid jobs

- The CCIPL will *not* become a grid site per se
- We need an entry point though :
  - we can put a VOBOX inside the CCIPL
    - that VOBOX would need to be able to do direct submission to the CCIPL batch system (which is not defined yet) : not a big issue I assume ?
    - other options ?
- As the site won't have storage, expect only MC productions running there : can this be guaranteed somehow ?

Network

- Here's the one « major » issue we've identified so far : outgoing connectivity of the workers
  - Not clear at the moment if the workers can have public IPs (like we have in Subatech)
  - Outgoing traffic (in addition to input) is by default closed by the University firewall. Openings made on request only.
- If we cannot offer outgoing traffic for all workers, what are the alternatives ?
  - is xrootd 4.1(2?) proxy feature the solution ? any written recipe on how to setup ?
  - simple NAT ?

Next steps

- Prepare first procurement (now-fall 2015)
- Test implementation ideas (either on the current CCIPL or even locally in dedicated test setup if needed)