ADC TIM wrap-up

ADC TIM September 2017

- few topical sessions
  - left out a lot of ADC topics to have the time to go deep in the chosen topics.

- Focus on
  - Compute Resource description
  - HPC
  - Singularity
  - Data Access: WAN vs LAN, direct I/O vs copy-to-scratch
  - Event Service and Event Streaming Service
  - Clouds

Compute resource description 1

- Why do we ask? What are the use-cases?
    - Production campaign resource planning where not all workloads are equal
        - Reprocessing, HI, Upgrade work needing himem
    - Resource usage optimization, folding in urgency, budget and cost
        - hi and lomem jobs share a node without wasting cores - cheap
        - run more himem and leave cores idle - costs
        - some sites can run himem at no cost - no way to know currently!
    - Need help to formalize a toy use-case and add more
- What is exposed already?
    - Glue2 schema - usually not accurate as HW retired and added
        - averages may be sufficient (for planning) - load into agis with additions for ATLAS
    - HTCondor sites publish ClassAd - additions are easy and may become ‘standards’
    - Pilot probes cpu, benchmark, could store scratch space and RAM (maybe already?)

Compute resource description 2

- More job characteristics passed through CE to batch for packing/optimization
  - in addition to RAM, walltime, corecount - Disk io, iointensity, DbLoad, cpu efficiency
- We need to analyze the data we have already
  - To understand what we can already say
- In line with the first steps we discussed in the Harvester Bern meeting now 9months ago (the WN map):
  - Who does what? We need engagement from site experts, from batch sys experts, even just to help us in understanding what we want.

Harvester

- Grid
  - Plugins close to commissioning
  - Information collector to come (see previous slides)
  - Development of pilot stream control with the highest priority in next months
- Cloud
  - Full production at CERN OpenStack cloud with HTCondor plugins
  - Other resources to come, e.g. full CERN and LRZ clouds
- HPC
  - Theta/ALCF : Commissioning with 65536 cores (1/4 of Theta). Full production in one month
  - NERSC : Planning to migrate to Harvester with mini-pilot by the end of year
  - Titan and BNL KNL will follow
  - Yoda + Jumbo jobs (event service) : Ultimate goal for the most optimal usage of HPC compute power
- Harvester Core
  - Advanced features being developed
  - Monitoring to be followed up with bigpandamon team

HPCs

- Allocations:
  - US - 200M cpu hours, growth foreseen till 2030
  - DE, CH - 60M cpu hours, possible growth
  - CN - 17M cpu hours
- Many scenarios for job execution, should consolidate to AES mode
  - Already used at SuperMUC
- I/O bottlenecks, various techniques to reduce stress on shared filesystem, the most promising is to use container fat images
  - NERSC recipe to be followed up and discussed with software deployment team
- Advanced techniques (eg cvmfs) explored by CSCS PizDaint using Cray technologies
  - To evaluate if HPC can replace Tier-2
- Harvester testing at Theta
- Software distribution needs consolidation, containers can provide a common solution

Singularity

- Deployment model agreed - start with container launched by ATLAS wrapper
  - Central control/switch
  - Big sites need to be tested as soon as possible (all Tier-1s + big Tier-2s, possibly Tier-0)
- More elaborated models (eg step execution) require further discussion
- Start with proactive deployment on grid sites to follow up WLCG recommendations
- Site configuration issues and problems
  - Partially understood
  - Starting with clean environment needs to be followed up
- Small images/trees to be used for grid
- Fat images with ATHENA pre-installed for HPCs
- Follow up on common distribution with software developers and deployment team
- Docker to follow up after singularity is production ready on most of the sites

WAN/LAN & IO

- Direct I/O :
  - athenaMP able to support direct I/O. Nothing needs to be done on the transform part
  - Will add direct I/O to DDM dashboard and evaluate which fraction of the files is read
  - Some workflow might only use direct I/O (e.g. premixing).
  - Need overview of sites that are good at direct I/O
- Identified areas where we can improve I/O for WAN/LAN
  - Roadmap defined involving FTS, DDM, pilot
- WMFS side :
  - Test of Athena direct access, new list_replicas
  - To avoid scheduled transfers, brokerage needs to be adapted to consider WAN copies
- LAN/WAN access :
  - Need to have a plan to test WAN access
  - Switch to task level vs queue level for LAN/WAN access
- Test in HC to evaluate LAN/WAN access foreseen

Event Service

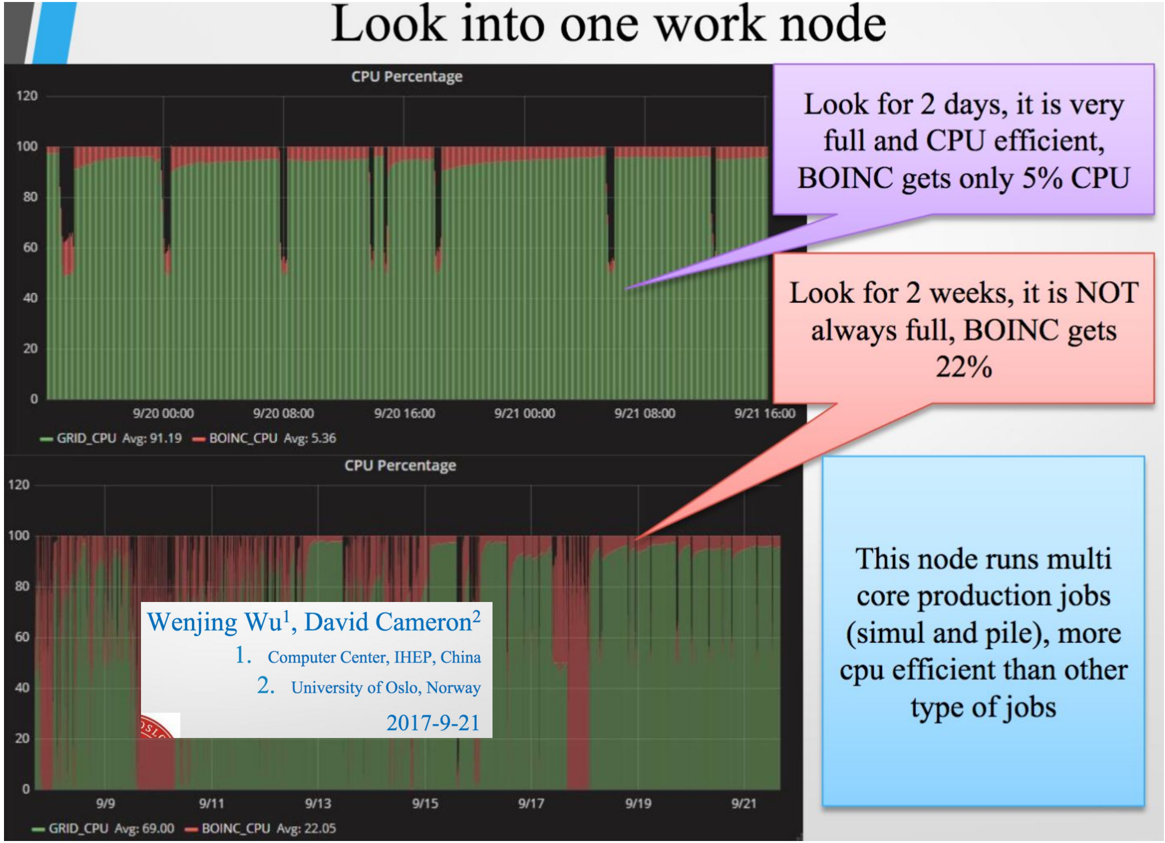
- In production
- Big progress in monitoring
- Validation: painful endless exercise
- Running for opportunistic and standard resources (with different upload freq)
- Yoda (Event Service on HPC) under active redesign and development
- Simulation remains the only workload supported by the ES. Need to start taking steps to extend it to other workloads. Derivation seems to be the first candidate
- ESS: challenging ideas, we need to start simple and evolve.
    - Simple prototypes exist. Need to complete the integration of these prototypes with Pilot
    - New manpower is available to start working on the first prototypes of the Server component (intelligent data delivery to the worker nodes)
    - We start with Simulation (the only workflow supported by the ES today)
    - We need deep study (analytics) to see where and what we can optimize

Clouds 1

- Review of cloud resources shows compute is stable, but without growth
- Objectstore use-cases continue to be studied
- Proposal to study other cloud services which may be useful to ADC
    - Cost model to be understood
- Dynafed discussion on Content-MD5 checksum for file transfers
    - Agreed to push usage of RFC2617 within data management tools for clouds, to be understood the impact on the rest of the Grid storages.
- Shoal development now provides auto-discovery of Frontier proxy discovery
    - Could this simplify some “lightweight site” config?
- Proposal to simplify sim@p1 provisioning:
    - investigate container tools for this
    - Careful interactions needed with TDAQ team

Clouds 2

- Running BOINC alongside existing jobs on T2 WNs was studied
- 22% extra CPU-hours can be extracted
- reduced priority of BOINC jobs allowing better scheduling within a WN
- no impact on existing job performance



Look into one work node

Look for 2 days, it is very full and CPU efficient, BOINC gets only 5% CPU

Look for 2 weeks, it is NOT always full, BOINC gets 22%

This node runs multi core production jobs (simul and pile), more cpu efficient than other type of jobs

Wenjing Wu[1], David Cameron[2]

1. Computer Center, IHEP, China
2. University of Oslo, Norway

2017-9-21

ADC (r)evolution 1

- Can’t summarize on the fly 2hours of presentations
- Looking backwards 10 years ago we see that we had no idea how now would have been.
  - We need to plan our frameworks to be flexible enough to cope with the future
- Cost of our resources, disk and storage
  - As soon as we (someone) will start paying for real, for sure we will be pushed strongly to optimize much more.
  - metrics!
- Automation, automation, automation.
  - Need cleanup first, need (sometimes) to break backwards compatibility
- .

ADC (r)evolution 2

- Evolve our frameworks towards other communities
  - We are not alone!
- The evolution can be very useful to us
  - Our framework will be more flexible
- This should be started within ADC
  - It’s not only copyright, it’s a matter of collaboration and coherency , thus stronger, in exposing and exporting our work.
  - We have to sit down and discuss.

ADC & SW … and the physicists

- We are not alone
  - We (ADC) run what SW guys gives to us
  - But we are all we inside ATLAS
  - Critical to feedback from ADC to SW the improvements “needed” and timeline.

June 2018 S&C Week

- As promised by Simone on Monday, we have news on an outside venue for the June 25-29 2018 S&C Week
- Thanks to David South and our DESY colleagues, we will hold the meeting at DESY
  - Details will come in the December S&C Week
- A big thank you to DavidS and DESY! An excellent venue for us.

Reminder -- this means no S&C meetings larger than 30 people outside CERN in 2018 unless we seek an exception to the rule. Smaller meetings are OK.

Summary

- Lively meeting
- A lot of discussions
  - Not always consensus
  - A lot of ideas ➜ == a lot of work!
- Analytics: one bit we did not tackle these 3 days
  - But mentioned *each* day. Maybe possible to get some ATLAS physicists to help?
- Early prototypes: yes!
  - And then review which one to keep, drop the others.
- Today is the right time to start (we actually never stopped) (re)thinking about (r)evolutions
  - Incremental when possible, but being prepared for jumps
- Communication/sharing info is paramount: suggestions are welcome!
  - (but we do not want even more meetings!)
- **People**
  - **We** (you) are the key!