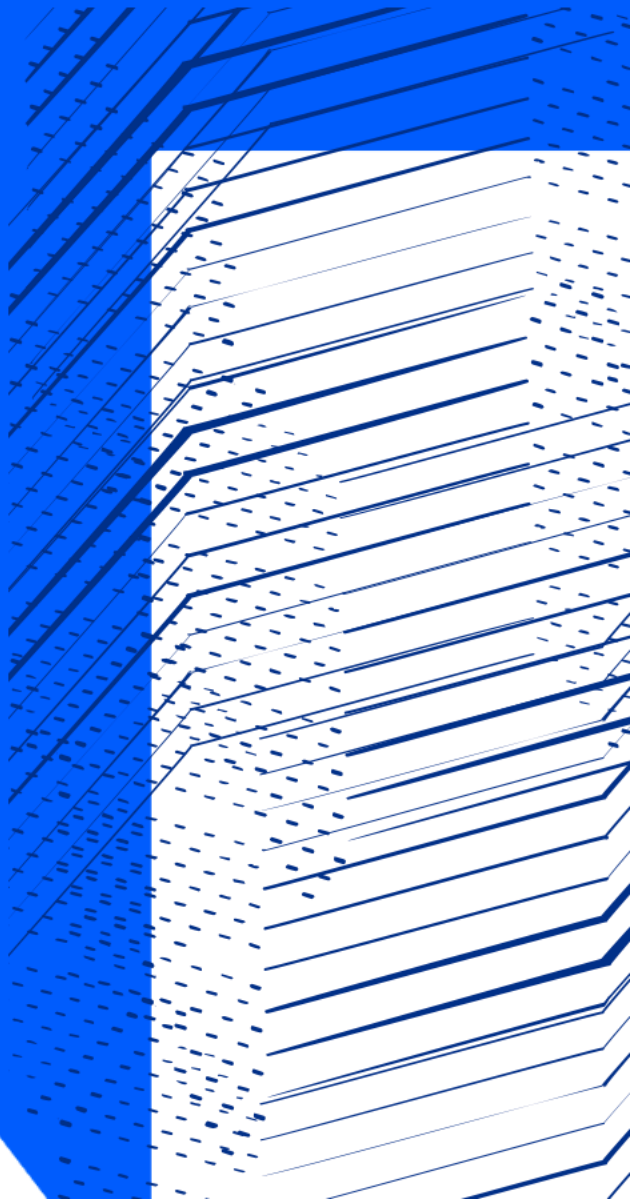
Science and Technology Facilities Council

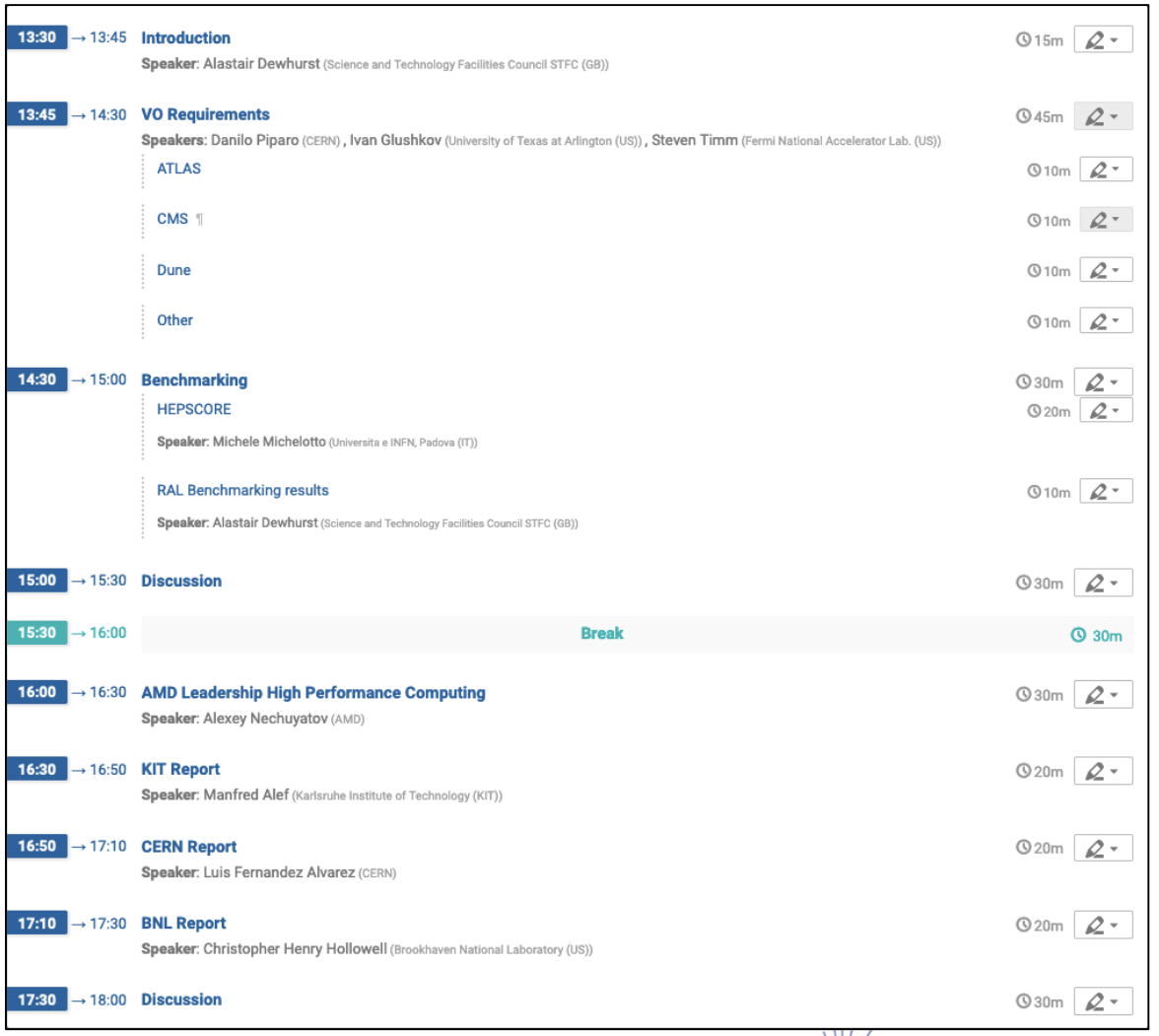
Pre-GDB on WN Summary

Alastair Dewhurst



Overview

- First half of meeting was attended by 40+, second half around 30.
- Thank to everyone who was flexible with the evolving agenda.
- Requirements
- Benchmarking
- Procurement
- Site Reports



13:30 → 13:45 **Introduction** 15m
Speaker: Alastair Dewhurst (Science and Technology Facilities Council STFC (GB))

13:45 → 14:30 **VO Requirements** 45m
Speakers: Danilo Piparo (CERN) , Ivan Glushkov (University of Texas at Arlington (US)) , Steven Timm (Fermi National Accelerator Lab. (US))
- ATLAS 10m
- CMS ¶ 10m
- Dune 10m
- Other 10m

14:30 → 15:00 **Benchmarking** 30m
- HEPSCORE 20m
  Speaker: Michele Michelotto (Universita e INFN, Padova (IT))
- RAL Benchmarking results 10m
  Speaker: Alastair Dewhurst (Science and Technology Facilities Council STFC (GB))

15:00 → 15:30 **Discussion** 30m

15:30 → 16:00 **Break** 30m

16:00 → 16:30 **AMD Leadership High Performance Computing** 30m
Speaker: Alexey Nechuyatov (AMD)

16:30 → 16:50 **KIT Report** 20m
Speaker: Manfred Alef (Karlsruhe Institute of Technology (KIT))

16:50 → 17:10 **CERN Report** 20m
Speaker: Luis Fernandez Alvarez (CERN)

17:10 → 17:30 **BNL Report** 20m
Speaker: Christopher Henry Hollowell (Brookhaven National Laboratory (US))

17:30 → 18:00 **Discussion** 30m

Requirements

- Talks from ATLAS and Dune.
- Other LHC VO had very similar requirements to ATLAS.
- Dune (and other non-LHC communities) had higher memory requirements.
- 8 core job slots seen as a ”sweet spot” for VOs.

ATLAS requirements (per score slot):

- **Mcore Slots:** 8-core is optimal (more cores, less efficiency)
- **HDD:** 20 GB (min. 10-15 GB)
- **RAM:** 2 GB (better: 3-4 GB)
- **Swap:** RAM+Swap >= 4GB
- **Network bandwidth:** min. 0.25 Gbit/s (CPU speed dependent)
- **Software environment:** OS, CVMFS, all jobs run in Singularity containers

Benchmarking

- Talk from Helge summarizing the work of the WLCG task force.
- Talk from Michele detailing the benchmarking framework.
- Talk from me with some benchmarking results.

HEP-SCORE Deployment Task Force

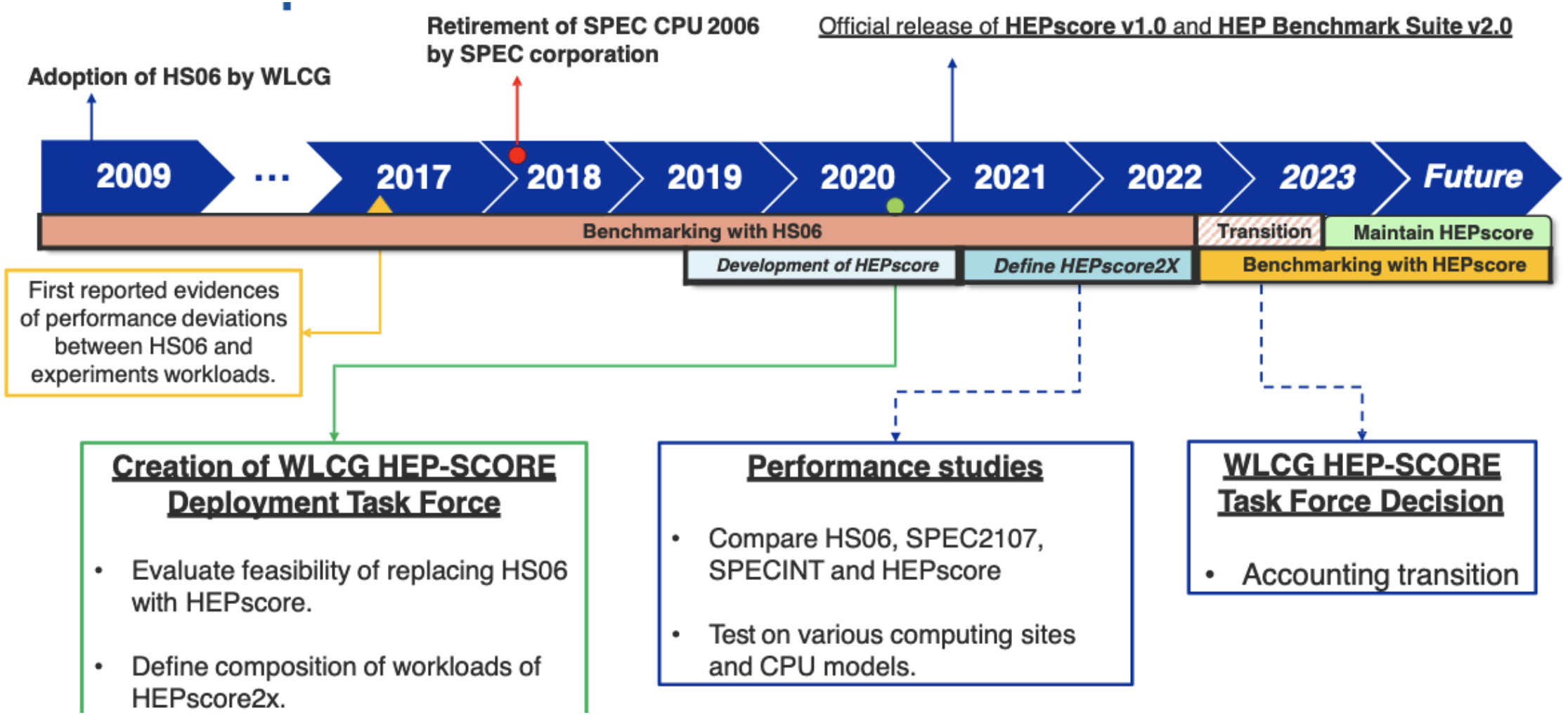
- WLCG Management Board discussed and decided to launch a task force
- Started in November 2020, bi-weekly meetings since then
- Membership:
  - Experts on pledge etc. process / procurements
  - Experiment experts
    - Four LHC experiments
    - Belle 2, DUNE, LIGO/Advanced VIRGO(/KAGRA), JUNO/BES III etc.
  - Site experts
  - Some MB members

Results - AMD

Dual SKU	Memory	HEPSCORE	SPEC2006	SPEC2017	Average Power (Watts)
75F3 (32C / 64T)	1024 GB	2492	2592	324	
7313 (16C / 32T)	512GB	1183	1299	167	431
	1024GB				463
7543 (32C / 64T)	512GB	2181	2328	300	577
	1024GB				615
7763 (64C / 128T)	512GB	3507	2855	395	674
	1024GB	3654	2938	401	693

- Variety of the Milan CPU tested.
- For reference, last years 7452 has 1800 HS06.

HPscore timeline



https://indico.cern.ch/event/876806/contributions/4400256/subcontributions/344688/attachments/2280880/3875385/HEPSCORE-WLCG-Michelotto-20210713.pdf

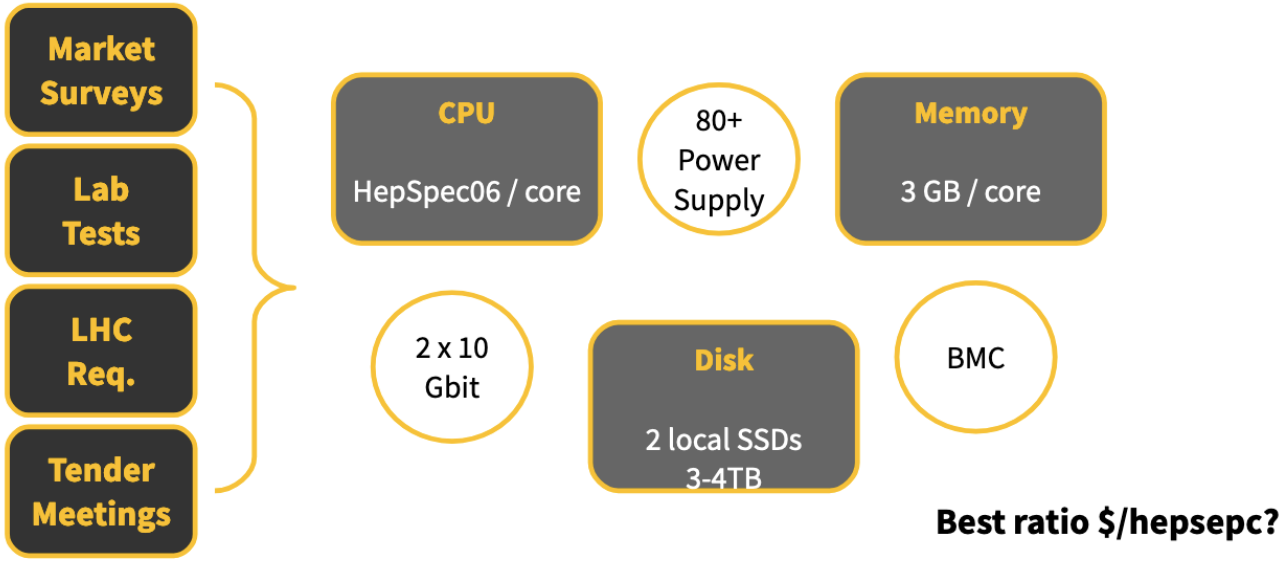
Site Reports

- 3 excellent talks from KIT, CERN and BNL
  - Provided details on WN specs, procurement and setup

Hardware Details

- Default setup: SMT on, ~1.5 job slots per physical core (multiple of 8 job slots to optimize 8-core job scheduling)
- Memory:
  - At least 4 GB RAM / (phys.) core
    ⇒ ~3 GB RAM / job slot in default setup
  - (Older WN models providing at least 2 GB RAM / job slot)
- Hard disks, SSD(s):
  - At least 30 GB scratch space and swap (> 4 GB) per job slot
  - Latest WN models: SSD(s), older systems: multiple hard disks
- Network connection: 10GbE (older WN models: GbE)

Tendering process



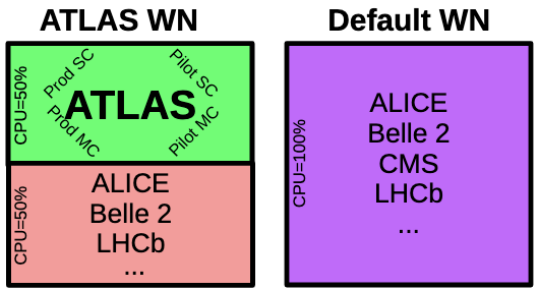
Procurement

- I have been very worried about procurement problems this year.
  - I even arranged for AMD to talk at the meeting because of concerns over supply problems.
  - AMD slides should be uploaded at some point.
- While it is certainly not a straight forward year for procurements, the general feeling was it is not so bad.
  - Delays are of the order of weeks (not months).
  - Memory is more expensive currently.
- Don’t leave it to the last minute, but there is no reason to panic.

Job Scheduling

- Several mentions of scheduling problems between single and 8 core jobs.
  - Primarily involving ATLAS because they run the most mixed workflows.
- There doesn't appear to be an optimal solution currently.
  - Sensible to pursue this further. E.g. HTCondor workshop.

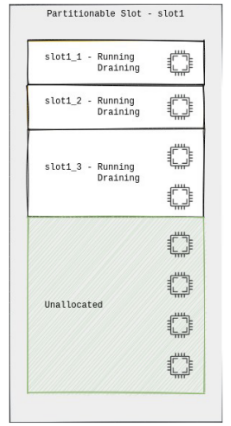
- Modified setup optimizing ATLAS unified SC/MC queue scheduling
  - Several issues caused by frequent MC-SC-MC transitions
  - Solution (since autumn 2020): dedicated ATLAS farm partition, still maintaining job mix by implementing 2 HTCondor partitionable slots on subset of WNs, serving either only ATLAS, or only the other VOs



Fragmentation

Our cluster runs a mix of single-core, multi-core (8) and arbitrary job sizes

- Fragmentation becomes a problem as the vast majority of the job requests are single-core
- How to find a fair allocation for multi-core jobs?
  - Current approach: condor_defrag. Drain nodes to allocate multi-core jobs.
  - Once a machine has at least 8 cores available, it only accepts multi-core jobs for a few negotiation cycles.
- Challenge to find the sweet spot of concurrent number of machines to defrag



Replacement to CentOS7

- At the end of June the UK GridPP community had a discussion on the replacement to CentOS7.

Potential Issues For Upcoming Purchases

- **Pandemic-related hardware supply issues, but our sources indicate that delays are not expected to be too lengthy (perhaps weeks, not months)**
  - Situation seems to be worse on the storage side
- **CentOS 8 Early EOL**
  - Adoption of CentOS/RHEL 8 for worker nodes has been delayed in the community while a decision on the path forward is made
  - Rocky Linux 8.4 was released in late June
    - BNL is currently testing, but this is likely the solution we will adopt for compute
  - AMD Milan CPUs not officially supported in RHEL/CentOS/SL 7
    - https://access.redhat.com/articles/5899941
    - Sites continuing to run RHEL/CentOS/SL 7 on bare metal wanting an officially supported configuration must continue to purchase AMD Rome, or Intel (Ice Lake or Cascade Lake) CPUs

Future OS

If we wish to buy the best available hardware and run supported Operating Systems on it (which we do!). We will need to pick a new base OS very soon.
Jobs can run in CentOS7 containers for as long as necessary.
We know that this is something CERN and Fermilab are working on jointly and we request they urgently consider making recommendations.

- Rocky Linux 8.4 was released in late June
  - BNL is currently testing, but this is likely the solution we will adopt for compute
- AMD Milan CPUs not officially supported in RHEL/CentOS/SL 7
  - https://access.redhat.com/articles/5899941
  - Sites continuing to run RHEL/CentOS/SL 7 on bare metal wanting an officially supported configuration must continue to purchase AMD Rome, or Intel (Ice Lake or Cascade Lake) CPUs

Science and
Technology
Facilities Council

Questions?