COSMA8:

DiRAC-3

ICC Theory Lunch
25th January 2021
Alastair Basden

News

- Downtime
  - Next week, starting 1st Feb
- New cosma-support member
  - Aqeeb Hussain, starting shortly

COSMA 8.0

- Ordered March 2020
  - Archer-2 mitigation system
  - Delivered Sept 2020
  - Production Oct 2020
- 32 nodes
  - 128 cores per node (AMD Rome)
    - 4096 cores
  - 1TB RAM per node
  - Infiniband 100Gbit/s (EDR)

COSMA 8.0

- 2x login nodes - 2TB RAM, 64 cores
- mad04 - 4TB RAM, 128 cores
- GPU servers:
  - gn001 - 10x NVIDIA V100 GPUs
  - ga001-3 - 6 AMD MI50 GPUs

DiRAC-3

- £20m awarded Oct 2020
  - £5.4m for the Memory Intensive system (COSMA)
    - Of which £0.6m was for “data curation”
- This will be used to extend COSMA 8:
  - i.e. COSMA 8.1

Procurement process

- ~2-3 month procurement process
  - Thorough benchmarking of options

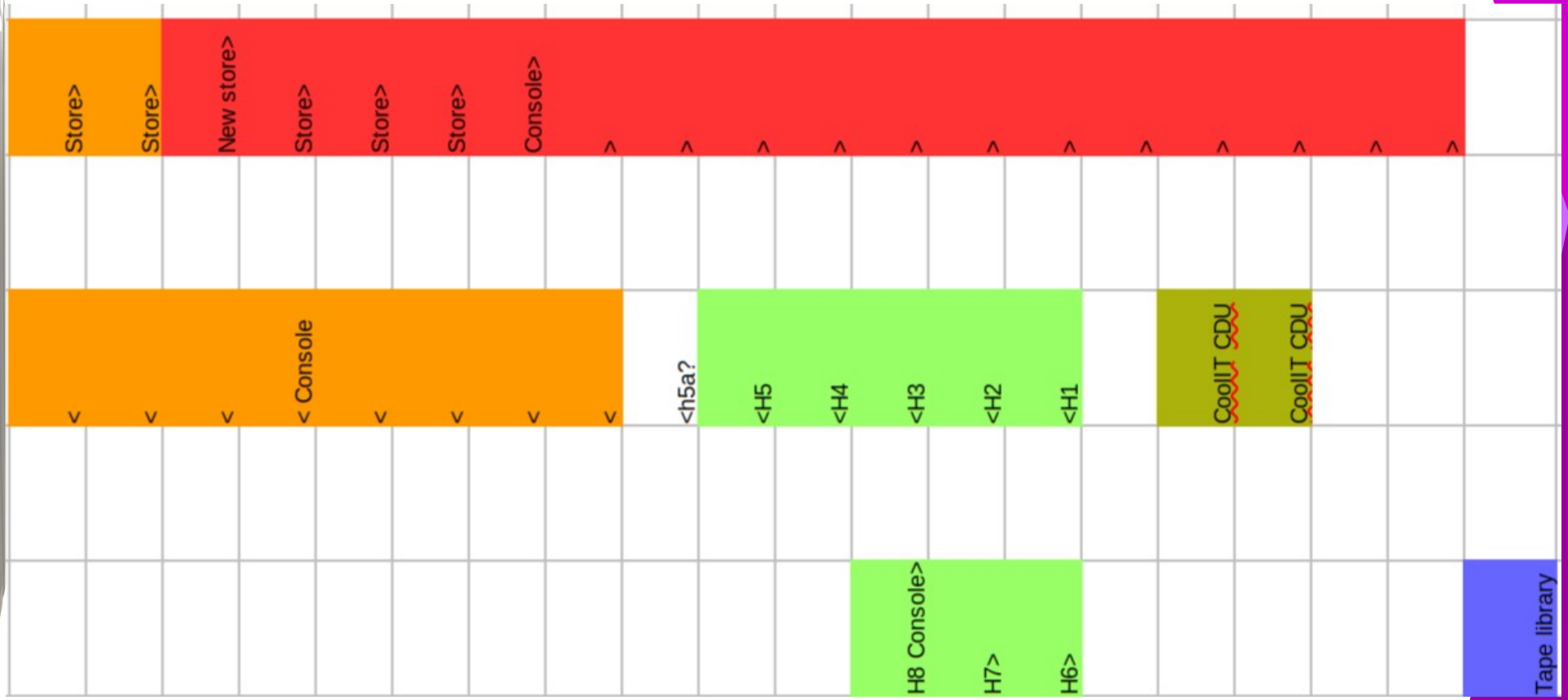
COSMA 8.1

- Ordered just before Christmas
  - Now beginning to arrive

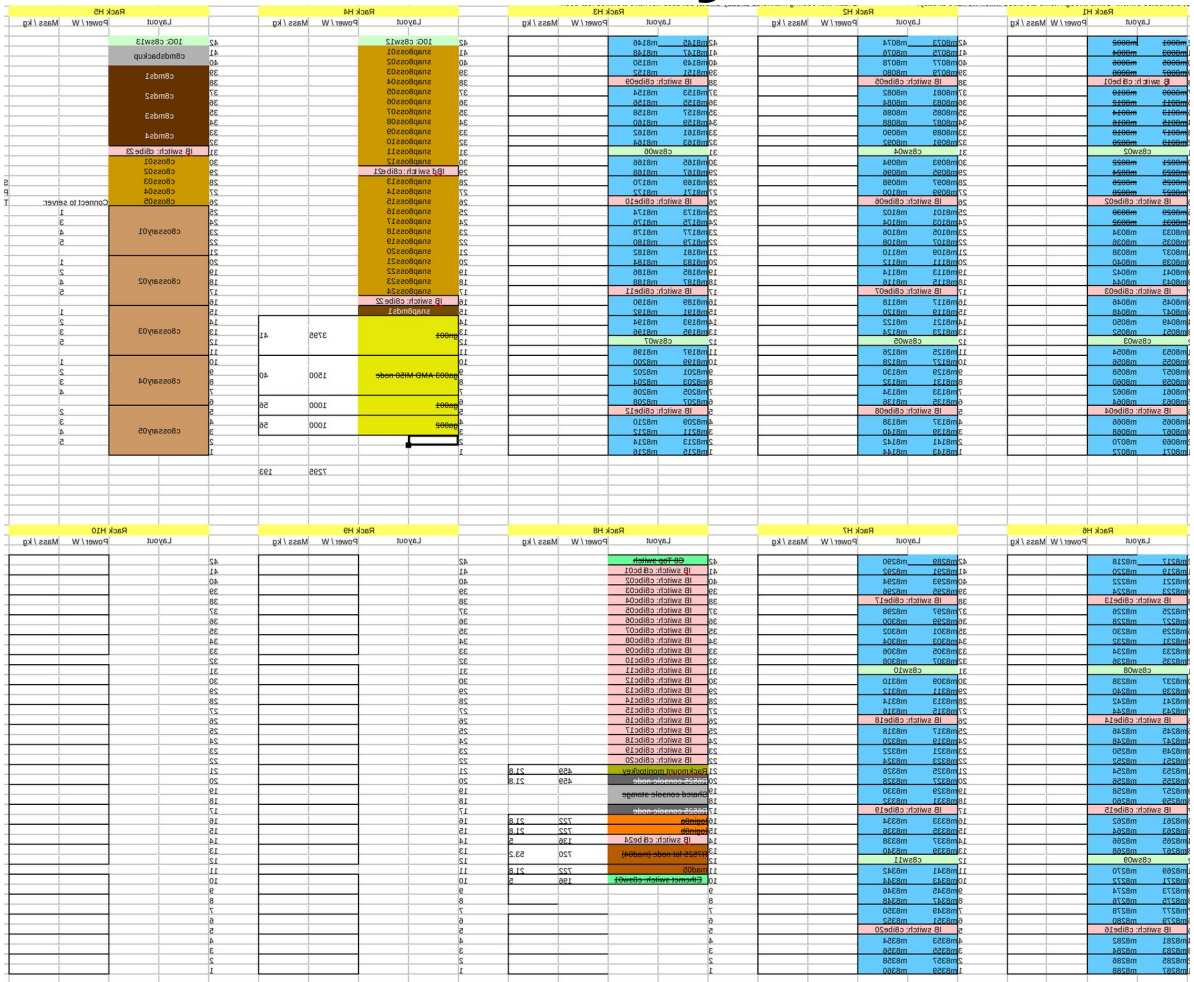
What do we get?

- 360 compute nodes (including the original 32)
  - 128 cores/node, 1TB RAM (£62/core full node, £7.75/GB, £20/core processor)
    - 46k cores (3.6x COSMA7)
    - 360TB RAM (1.6x COSMA7)
  - HDR200 Infiniband fabric
  - 4.5PB Lustre storage (/cosma8)
  - 1.2PB fast scratch storage (/snap8)
  - 2x fat servers (high memory, 4TB, 128 cores)
    - mad04, mad05
  - 2 login nodes
  - Redundant tape library (in OCW)
  - GPU servers

Where?


Store>
Store>
New store>
Store>
Store>
Store>
Console>
< Console
<h5a?
<H5
<H4
<H3
<H2
<H1
CoolIT CDU
CoolIT CDU
H8 Console>
H7>
H6>
Tape library

Rack layout



Things of note

- Direct liquid cooling (on-chip water cooling)
  - Processors are 280W each
  - First such installation in Europe
- A high RAM system (as before)
  - Reduced to 8GB per core

/snap8

- Should offer >2x performance of /snap7
- No redundancy
  - This is not a general file store
  - Please leave it for restart files, and remove anything not required
- Cost ~ £540k
  - 68 compute nodes
  - 8700 cores

Software stack

- Carried over from COSMA7
  - Some additions, some recommendations
- AMD processors: x86 compatible
  - Main difference: no AVX512
    - Will not run code optimised on COSMA7

Compiler recommendations

- Intel compilers probably provide best performance
  - use the latest version
- The AMD AOCC compiler
  - Based on llvm
- gcc - use version 10 or higher

Math libraries

- Intel MKL (Math Kernel Library) known to be hobbled
  - There are workarounds on the COSMA support pages
- Alternatives:
  - openblas
  - AMD libm

Notes

- With 128 cores/node, codes need to be well optimised
  - Please spend some time making sure you are using the nodes efficiently
  - If not, please use COSMA7 instead
  - Think about how to use it
    - We can no longer afford to run codes blindly
      - e.g.: 10% computation single threaded, 90% multi-threaded
      - On COSMA7 (28 cores), 10 seconds + 3.2 seconds
      - On COSMA8, 10 seconds + 0.7 seconds
- Nodes will eventually be non-exclusive:
  - If you ask for 1 core, you will get 1 core and 1/128TB RAM
    - Sharing the node with others
  - Requests for >=128 cores will give whole numbers of nodes

Access

- Operational late April (we hope)
- Released to DiRAC RAC in October
  - ICC get 20% of time (via dp004)
    - Plus the DiRAC RAC allocation
- Between then, available for testing for ICC
  - After initial run-in and acceptance testing
  - Ideal time for large jobs
  - Please start planning

Dell Centre of Excellence

- ICC will become one of these
  - Highly prestigious
  - Focused around future network and RAM technologies
  - Very open to new ideas for codes
    - Gen-Z (shared memory)
    - Switchless hypercube fabrics (consistent low latency)
    - Data processing units (programmable network cards)

COSMA 8.2

- Additional £2.6m expected in 2021
  - Could deliver an additional ~240 nodes
  - And 4.5PB storage
- Also happy to add other sources of funding
  - e.g. the recent departmental call for PDRA funding

Other facilities

- Bluefield cluster
  - 16 nodes
    - 32 cores, 512GB
    - BlueField data processing units
    - Switchless network fabric
- 3x A100 GPUs
  - In a Liqid infrastructure

Conclusion

- Start planning big…
- Change of date for Theory Lunches
  - 5 weeks until next one, then every 4 weeks
- Topics for future talks...