Oxford ARC Provision Experience of Jade 2.5

ACIT-Hub Training CoDesign Day 2025 – Bristol

Gregory J. L. Tourte & The ARC Team

https://www.arc.ox.ac.uk

https://www.it.ox.ac.uk

Outline

1. Introduction

2. JADE 2.5 Technology Pilot
   - JADE 2.5
   - Hardware
   - Setup
   - Functional tests
   - Delays
   - Operational issues

About Us

- Advanced Research Computing team at the University of Oxford;
- Part of central IT services;
- Provide generic, non-subject specific, high performance computing facilities for the whole institution;
- Provide help with software installation and minimal support for use on the cluster;
- Other facilities for more subject-specific uses at departmental/division levels are also available;
- About 1300 active users, across 600 active projects;
- Provide training to HPC users from introduction to advanced uses of HPC facilities.

Oxford ARC compute estate

- ARC capability cluster (14 640 CPU cores):
  - 305× 48 core worker nodes;
  - 2× Intel Platinum 8628 24 core 2.90 GHz Cascade Lake CPUs;
  - 384 GB memory;
  - HDR 100 InfiniBand interconnect (The fabric has a 3:1 blocking factor with non-blocking islands of 44 nodes);
- HTC high throughput cluster:
  - 95× worker nodes;
  - Including 49× GPGPU nodes;
  - 2× high memory (3 TB) nodes;
  - About 20 nodes with HDR 100 interconnect;
  - 10 Gbit Ethernet;

Oxford ARC compute estate

- JADE 2.5 technology pilot cluster:
  - 3× 128 core worker nodes;
  - 2× AMD EPYC 9534 64-Core CPUs and;
  - 8× AMD MI300X GPUs;
  - 2× NDR 200 networking;
  - 32 TB NVMe SSD local scratch.

JADE 2.5 Technology Pilot

Oxford ARC storage estate

- Lenovo OnTap storage
  - NFS shared large capacity storage;
  - Project and user home areas;
  - Software repositories;
  - 4.9 PB raw (3 PiB useable) total capacity;
  - 4× 25 Gbit/s aggregated interfaces for data traffic to clients.
- Weka storage
  - NVMe based storage;
  - Ultra low latency;
  - Used for scratch, databases, large shared datasets;
  - Total of ~700 TB of usable NVMe SSD;
  - Filesystem exported natively over both Ethernet and InfiniBand.

Training and Support

- Number of courses run regularly throughout the year;
- Online drop-in sessions available between courses;
- Extensive online documentation for the systems;
- Every day support via ticketing system.

Provisions Stories

- Several provision processes carried out over the last couple of years;
- Complete storage upgrade;
- new islands (partial cluster upgrades);
- JADE 2.5 Technology pilot.

Outline

1 Introduction

2 JADE 2.5 Technology Pilot

- JADE 2.5
- Hardware
- Setup
- Functional tests
- Delays
- Operational issues

JADE 2.5

- Technology Pilot
- Based on 24x AMD MI300 GPUs linked by infinity fabric.
- Small system — three servers (8 GPUs each).
- Aim to start onboarding users in November.
- Time will be allocated — about a month of time per partner.
- Plan to also schedule ‘student weeks’.
- Primary objective is to asses the AMD GPUs’ suitability for this type of research, rather than directly conducting research.
- Project allocations come with a requirement for feedback on system performance and behaviour.

Hardware

3 Lenovo ThinkSystem SR685a with:

- AMD MI300X GPUs with 192 GB GPU memory
- 8 GPU Board, linked by Infinity fabric
- 2x AMD EPYC 9534 (128 cores per server)
- 2.25 TB RAM
- 32 TB node-local ‘scratch’ storage
- 960 GB M.2 (RAID 1) for OS
- Networking:
  - NVIDIA ConnectX-7 NDR 200 IB
  - Mellanox ConnectX-6 Lx 25 GbE
- 8x 2600 W power supplies (2N redundancy)

Hardware

Hardware

- 8U server; weighting approximately 110 kg
- Requires 1200 mm racks.
- Estimated power draw per server:
  - 6715 W at 85 % workload
  - 7946 W at 100 %

Notable installation — first systems of this type installed in the UK.

Hold that thought.

Setup

- Systems built using ARCs existing stack (PXE, kickstart, CFEngine).
  - No significant issues during installation.
- Integrated into existing ARC infrastructure (LDAP,...).
  - Dedicated SLURM controller and SLURM database.
- Shares ARC networking and storage infrastructure:
  - HDR fabric (interconnect, fast storage).
  - WEKA file system for non-local scratch.
  - Lenovo DM3010H for ‘bulk’ data storage.
- Software environment:
  - ROCm installed via RPM (using AMD’s repositories).
  - Managed with EasyBuild — minimal set of pre-installed toolchains.
  - Users are expected to bring (or build) their own software.

Functional tests

- ROCm Validation Suite:
  - Testing caused node crashed (more details later).
- TensorFlow container:
  - Successfully run on system — no issues observed.

Delays

- ‘Equipment must be installed and invoiced by 31st July 2024’.
  - Quotes received in May; system ordered early June.
  - System delivered 11th September 2024.
- RFQ specified InfiniBand (HDR fabric).
  - SR685a initially only supported with 4× Broadcom Ethernet cards.
  - Qualification of the systems with IB cards completed just prior to delivery.
  - Systems shipped with Broadcom cards installed.
  - NDR 200 cards arriving separately.
  - Required the ARC team to swap the cards.

Delays

- After installation, Lenovo identified that the as built power distribution board could not adequately supply power to the GPUs with the latest AMD firmware.
  - Performance reduction of 5 % to 10 %
  - Required replacement of power distribution board.

Operational issues

- Operating System Compatibility — requirements stated ‘Nodes must support any Linux distribution ABI compatible with Red Hat Enterprise 8 and/or 9, and should ideally also support Ubuntu LTS.’
  - When shipped, the platform only supported Ubuntu 22.04 LTS.
  - Red Hat Enterprise 9.4 now supported (installed OS: AlmaLinux 9.4).
- ROCm installation/ROCm compatibility:
  - Installed latest ROCm.
  - Initial testing using ROCm Validation Suite were promising.
  - Running MI300X test configurations caused severe system crashes.
  - Setting a power cap helped, suggesting the PDB issue as cause.
  - However, response from AMD states that the systems are not currently supported by latest ROCm and indicates this as root cause of these crashed.

Conclusions

- Successful set up — no issues getting the GPUs to work.
- Ran test workloads almost immediately after installation.
- Significantly delayed delivery.
- Minor issues encountered throughout setup, likely due to the newness of the system — but these add up, impacting overall perception of the platform stability.
- The platform requires a very specific combination of firmware and driver versions. This was/is not clearly indicated by AMD's ROCm compatibility matrix — improvements are needed in the published information.

Conclusions

- Successful set up — no issues getting the GPUs to work.
- Ran test workloads almost immediately after installation.
- Significantly delayed delivery.
- Minor issues encountered throughout setup, likely due to the newness of the system — but these add up, impacting overall perception of the platform stability.
- The platform requires a very specific combination of firmware and driver versions. This was/is not clearly indicated by AMD’s ROCm compatibility matrix — improvements are needed in the published information.
- Several months on though...
- System is now very stable
- Very much in active use with so far 24 users from 5 Institutions

Thank You!

Any Questions?