A foray into composable infrastructure for HPC

Alastair Basden, Peter Draper, Paul Walker, Mark Lovell, Richard Regan,Gokmen Kilic

DiRAC / Durham University
ExCALIBUR H&ES

CIUK 2023

DiRAC

- UK national HPC service for STFC researchers
  - Tier-1 facility
- 4 sites:
  - Extreme Scaling @ Edinburgh
  - Data Intensive @ Leicester and Cambridge
  - Memory Intensive @ Durham
- Bespoke systems for the associated science
  - More cost effective than a single large system
  - Focus on Capability systems
    - For pushing the boundaries of what can be achieved

DiRAC

High Performance
Computing Facility

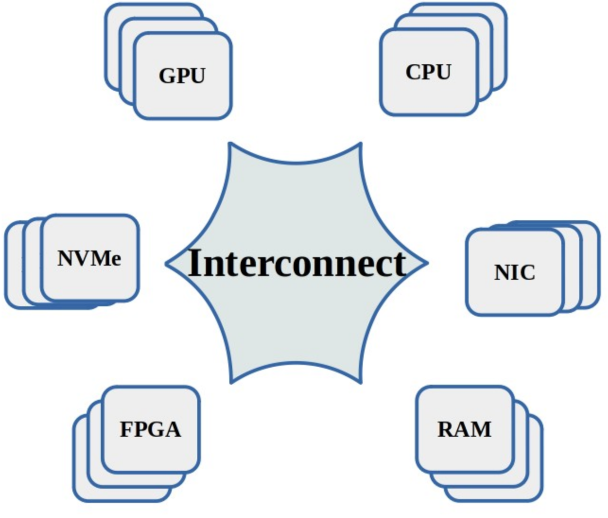
COSMA

- COSMA7: 452 compute nodes (115kW total)
  - 28 cores, 512GB RAM (~95kW)
  - EDR InfiniBand (100Gb/s) and Rockport 100Gb/s (6kW)
  - Fat tree 2:1 blocking
  - 6PB storage, 420TB fast NVMe (15kW)
- COSMA8: 528 compute nodes (~300kW total)
  - ~70k cores (~250kW)
  - 128 cores, 1TB RAM per node (Rome/Milan)
  - HDR InfiniBand (200Gb/s) (~18kW)
    - Fat tree non-blocking
  - 15PB storage (20kW)
  - 1.2PB fast NVMe storage ~350GB/s (8kW)
  - Cooling distribution units (1kW)

COSMA science

- Primarily cosmology:
  - Simulation of the universe
- Also nuclear physics, particle physics, black holes, planetary collisions, galaxy formation

Composability



- Separation of device resources
  - Compute, accelerators, storage, network, RAM
  - Treated as services
- Physical components no longer in a server
  - Assigned to the server upon demand
    - Building compute capability as required
    - Clusters can be better matched to typical use cases
- Dynamically provision bare metal via software

Compostability

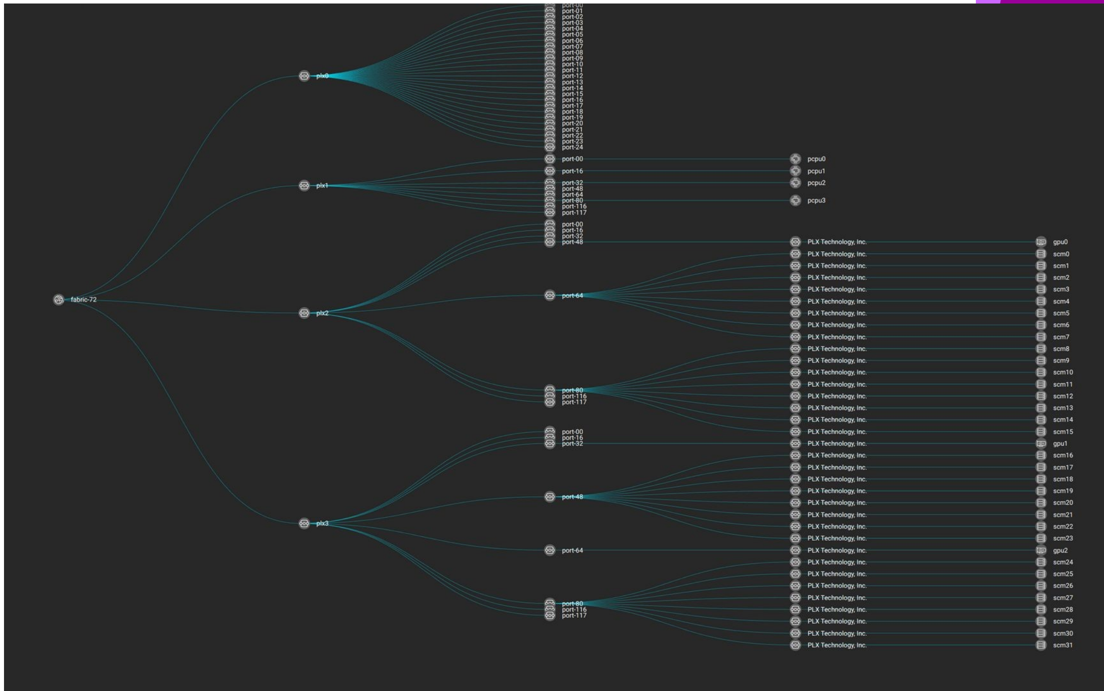
- Compostable infrastructure used for CIUK student cluster competition
- Not to be confused with composability
  - Similar aims (lower emboddied CO2, better resource use, etc)

Uses for composability in HPC/AI

- Massive GPU systems
  - 10s of GPUs per server
- Scarce resource sharing
  - Move GPUs as required to assigned servers
- Memory bursting
  - Adding RAM to servers as required
- Networking?
  - Composing multiple BlueField+GPU cards

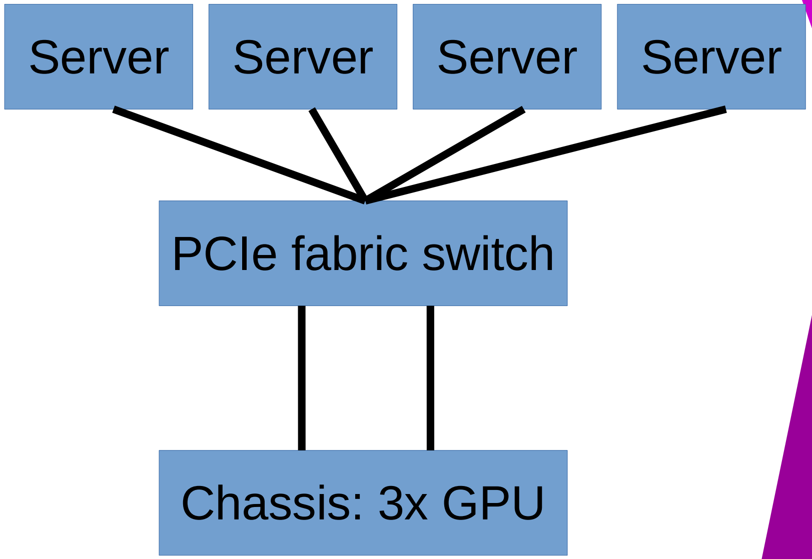
Composability on COSMA

- Liqid fabric:
  - One OSS chassis
    - 3 A100 GPUs (2021)
    - 4 RAM cards (3TB each) (2023)
  - PCIe switch and controller
  - 4 servers with fabric cards
    - One login node
    - 3 in a Slurm partition



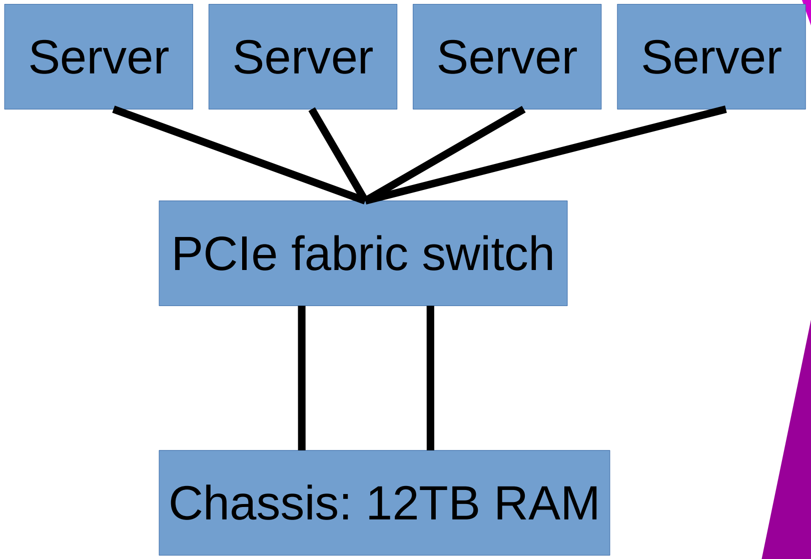
Liqid GPU system

- 3x A100 GPUs assignable to 4 servers
  - Can be hot-swapped
  - Occasionally causes problems
    - 3.10 kernel
- Usually static
  - One is a login node
- Slurm integration possible
  - Automatic provision
  - We use manual approach
- PCIe4 x4 connectors in each server
  - GPU bandwidth limited
  - 2 GPUs share 1 chassis card
- Physical connectivity is a pain
  - 4x SAS-type cables per card



Liqid RAM system

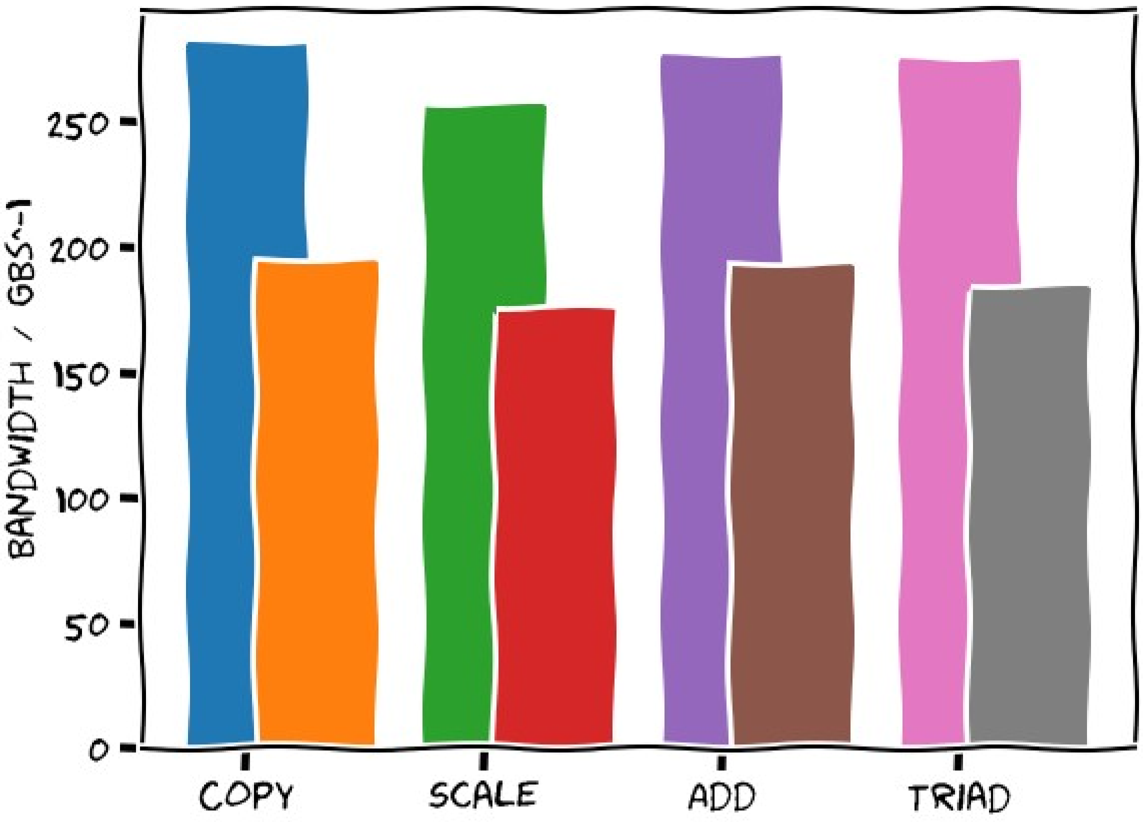
- 4x Honey Badger PCIe cards
  - Each with 8 M.2 NVMe Optane drives
    - 360GB - ~3TB total
  - 11.5TB combined Optane RAM
  - Each drive can be assigned individually to any server
  - Default configuration 8 drives/server (2 per HB)
- Memverge software to map Optane to RAM
  - Tiering of native RAM and Optane
    - Hot data kept in RAM, warm data moved to Optane
- Default setup adds ~2.5TB to each server
  - A bit of a pain to re-compose
- COSMA Jupyter hub can run on it
- Problems:
  - Lack of PCIe device entries in the BIOSs
  - Possible kernel bug
  - Power demands (native)



Memory bandwidth

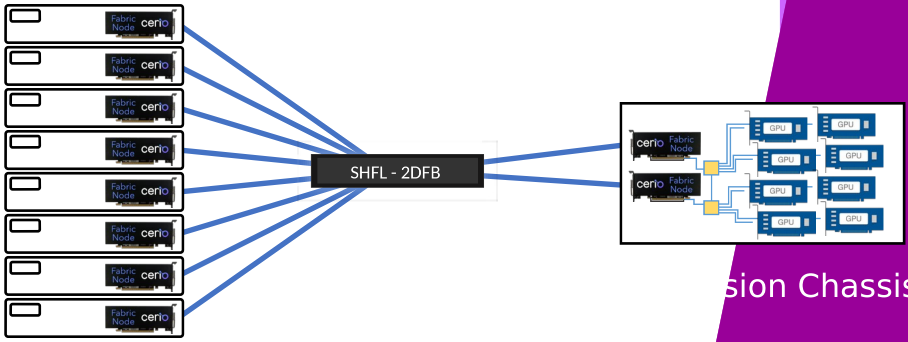
- Limited by the PCIe bandwidth (64Gb/s)
- MemVerge software used
  - tiered memory transfers between local DRAM and composed memory
  - Behind the scenes
- Memory allocations using a LD_PRELOAD
  - Prefix commands with “mm”
    - mm free -h, mm python jupyterlab, mpirun mm …, etc
    - Could be made the default option

STREAM benchmarks tests



Rockport Cerio system

- Moving from in-rack-scale to cluster-scale
  - PCIe-based fabrics don’t scale well
- Cerio flit-based fabric (6D torus) scales to thousands of nodes
  - Experience with the Rockport Ethernet fabric (COSMA7)
- High-density optical cables
  - Standard MTP connections
- Active components solely in server cards
  - “switch” not an active component
- 8-node test system in planning stage
  - 300GBit/s bandwidth, 200Gb/s to nodes
  - PCIe5

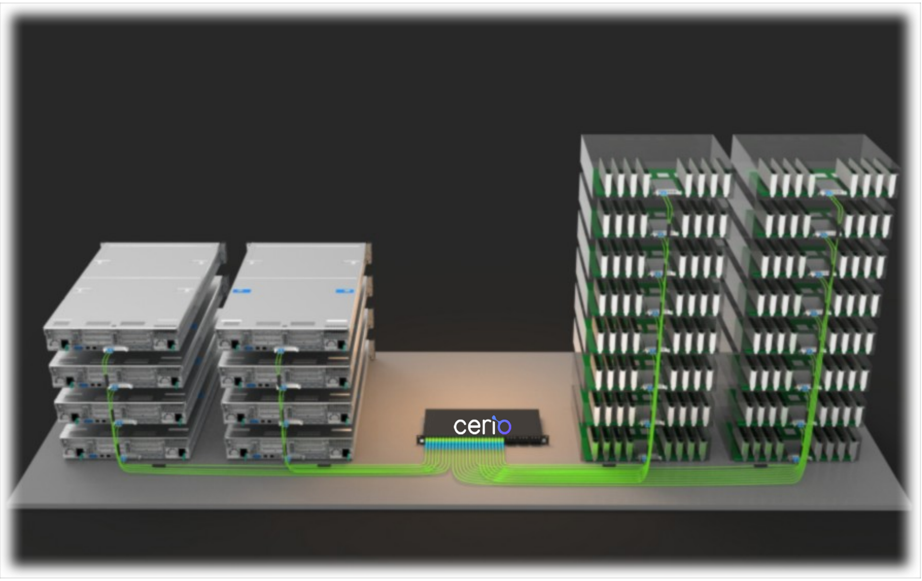


CXL and composability

- Compute eXpress Link
  - Open standard CPU to device and memory connections
- Brings memory coherence
  - Full composability of memory
  - Memory-coherent IO networking
    - Low latency communication - sub-mircosecond
  - /dev/shm spanning multiple servers
    - Multiple hosts work on same data without copying and shuffling

Considerations for composability

- Fabric reboots
  - Single point of failure
  - Cerio system should mitigate this to some extent
- Limited PCIe register slots
  - Determined by the BIOS
- Reliability
  - Particularly for RAM
- Cost: Not necessarily cheaper
- Power usage
  - Server thinks it is powering the cards
  - Custom firmware may be required
- Keeping track of infrastructure
  - Suspected dodgy components
- Not dynamically recoverable
- Bandwidth and latency hits
- Early days!



Net-zero considerations

- Significant potential
  - Lower embodied CO2 (less hardware)
  - Better match supply to demand
    - Less resource sitting idle
  - Expandable upon demand
    - Easy to add GPU resource as required by changing workloads

Conclusions

- Composability works
  - Currently a bit rough at the edges
- Lack of standards and flexibility
  - Vendor lock-in
  - Will hopefully improve
- RAM-based fabric could see significant performance improvement for some codes
  - CXL