Stepping towards HPC Cloud Bursting

Cliff Addison, Wil Mayers, Cristin Merritt and Manhui Wang

UNIVERSITY OF LIVERPOOL

Overview

- Liverpool in 2018 demonstrated strategic benefits of cloud for research.
- Deployment was ad hoc, but worked.
- We want to embed cloud for research and graduate teaching.

- 2019 – data centre move prompted focus on HPC resilience; 2020 hopes to expand on that

Successful cloud scenarios

- High throughput workflows
  - Current Windows Condor pool limited to circa 8 hr jobs
- HPC resilience
- Cloud bursting – more cycles needed for a short period
  - typically for papers or presentations
- Scoping studies
  - I think I need X cores and Y GB of memory for my research
- GPU nodes for Deep Learning

Cloud bursting - 1

- An existing Condor pool can be extended easily to the cloud.
  - Users just request the cloud resource on local Condor server – acts as scheduler.
  - Customise a standard AWS Linux image with necessary extra software and then save this image so is ready to go.
  - Have an in-cloud manager that deploys compute images; liaison with scheduler.
  - Spot market makes the compute even more cost-effective
    - Fits perfectly with Condor cycle stealing idea
  - Manchester have had good success here!

- Test that target instances are good enough

  Micro instances may be too slow so more expensive for compute

How this can work…

- Researcher came to us in May 2018 with an urgent request to run 100,000 simulations related to a paper under review.
  - Our AWS Condor pool ideal.
  - Cost per simulation cheapest on t2.medium, but fastest on c4.large or c5.large.

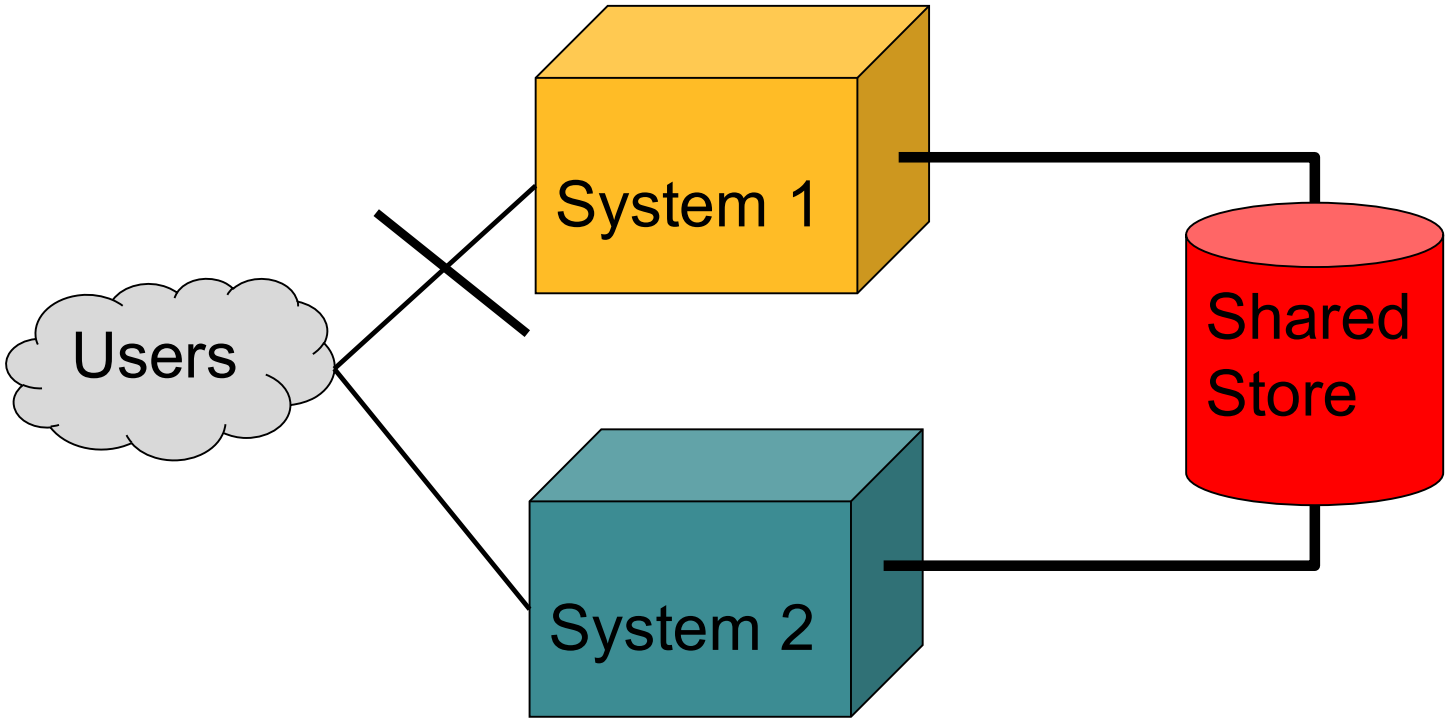
- Final set up:
  - 1000 jobs with 100 simulations each, pool size of 400 (so 400 jobs at once) completed task in **7h 21m**. Serially would need about **98 days** – massive speed-up.
  - Price **$51.16**

Paper resubmitted on time!

Moving forward on HPC front

- Have some basics in both AWS and Azure tenancies
  - Have Active Directory authentication for Azure
  - Direct Connect (faster network connections) still coming
- Liverpool HPC has:
  - Defined set of users (circa 80 active every month)
  - Stable software offering (slowly growing)
  - Alces Flight provide system and software framework
- Liverpool HPC needs:
  - Better resiliency – lacks failover component
  - More flexible environment for new users
  - Better development / experimentation support

Classical Active-Passive Failover



Issues with HPC resiliency

- Hard to support two on-campus HPC systems with an active-active failover mode (active-passive is silly)
  - HPC systems often sited in a single data centre
  - HPC systems bought at different times, maybe from different vendors
- HPC usage composed of many jobs running for hours, possibly days. [Not transactional]
- HPC storage geared towards performance and supporting a large number of simultaneous accesses
  - Hard (impractical?) to mirror all storage
- Not possible to migrate running or queued jobs.
- Cannot failover for brief outages.

What is mission critical for resilient HPC?

- Basic login and compute node environments.
- User authentication and authorisation.
- Mechanisms to load application environments and to submit jobs.
- Non-volatile user storage?
- Some compute nodes.
  - Replication of important node families, e.g. some GPUs
  - Interconnect for capability jobs (e.g. InfiniBand)?
- Budget for all of this??
  - Likely need to keep under control!

Scenarios where HPC resiliency relevant

- Power cuts to part / all of a data centre.
  - Power blips need to treated by other means
- Cooling infrastructure failure.
- Storage issues.
- Planned and prolonged system maintenance.
- Relocation or “swapping” HPC systems
- Need mechanisms that can kick-in automatically.

HPC resiliency in the cloud

- Want to have an on-demand clone available
- Compute can be brought on-line relatively quickly.
- Front-end / login node and storage need to be there through the life-time on the cluster.
- Compute node costs can be controlled via autoscaling options and by exploiting the spot-market (on AWS) – how many nodes are needed?
- Pricing of cloud compute for resiliency is an issue.
  - Most cloud platforms want a year of always on use before offering major discounts over their on-demand price.
- How deal with storage??

HPC cloud resiliency – storage issues

- Three types of data storage to consider
  - System – node images and applications
    - Persistent, relatively stable, modestly sized
    - Probably current to within a week is fine
  - User home directories
    - Many systems keep to a small number of TBytes for local backup
    - Daily incremental back-up to the cloud with occasional full back-up should be possible.
  - User volatile / work areas
    - These can be huge.
    - If shutdown is planned, can get relevant users to pre-stage important data; typically during the local rundown before shutdown..
    - Cloud as a primary and permanent site for volatile data?
      - Will be slow and might be very expensive…

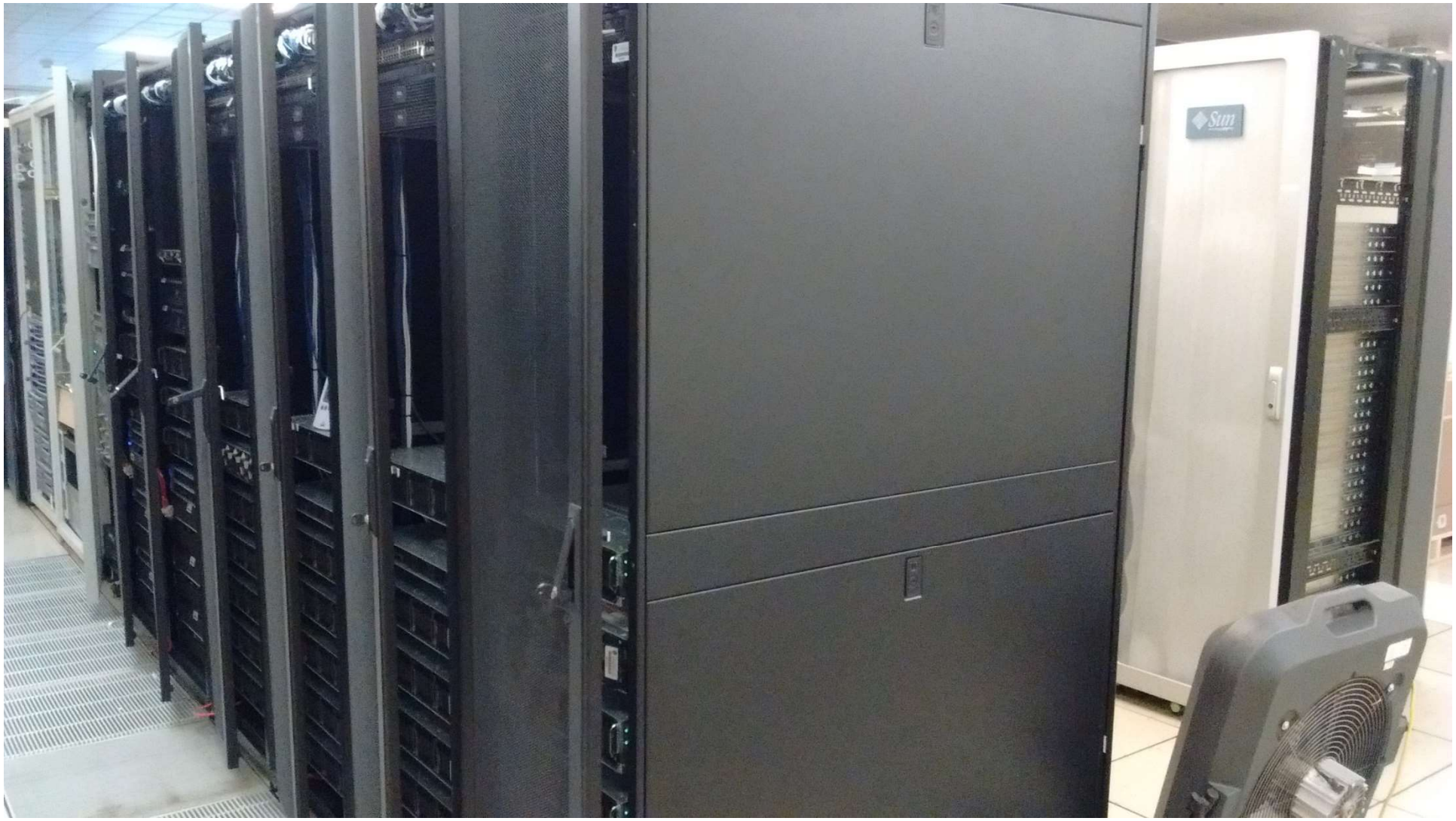
My understanding of AWS storage

- There are the traditional 3 storage layers:
  - Elastic Block Storage – fast access for active storage tied to hardware instances. Always need some of this on cluster. [100 GB costs about **$8.10** per month]
  - Standard S3 Object Storage – slower but accessible from anywhere in the AWS cloud (and elsewhere with S3 supported logical devices) [100 GB costs about **$2.30** per month]
  - S3 Intelligent Tiering, S3 Standard Infrequent access – slowly changing; not often accessed [100 GB/month **$2.40**, **$1.31** resp.]
- Also there are the archival options, not for HPC(?)
  - S3 Glacier and S3 Glacier Deep Archive – 6 month no-change?
- Other cloud vendors have similar arrangements.

2019 Motivation

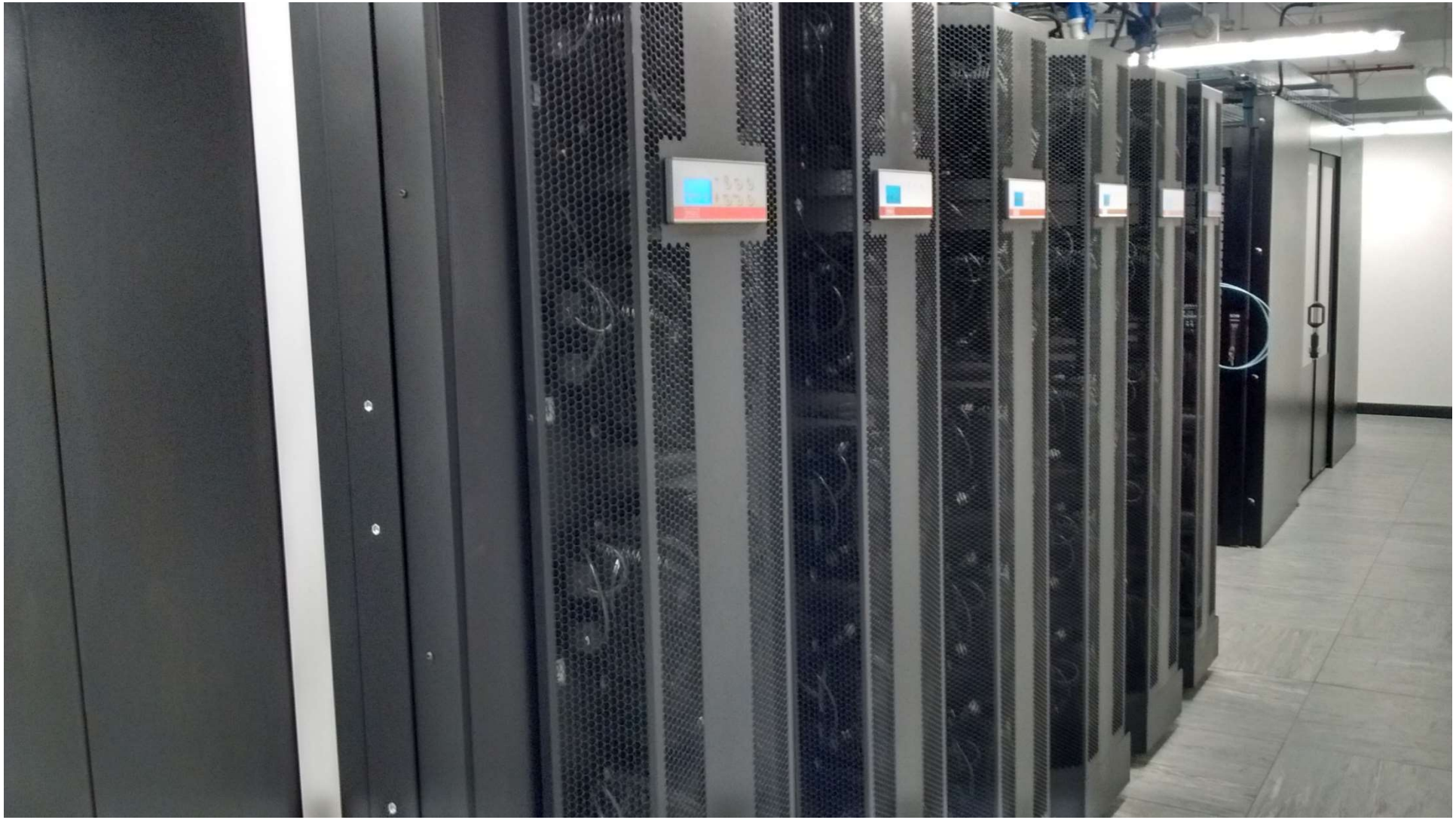
- New data centre with better cooling and generator-backed power for all systems finally finished.
- Needed to move Dell / Alces system to new home.
- Old Bull (SandyBridge) cluster available during this time, but lose 4000 cores for circa 10 days.
- Idea – augment SandyBridge cluster with some cloud-based Cascade Lake (AVX-512 support) and AMD nodes – great general purpose + GPU

September 2019 – 6 racks air-cooled



Racks generally look like this – cabling and storage challenge

October - moved to these water cooled racks



Trial run – June 2019

- Data centre 24 hr outage to swap power supplies
- Used outage to test cloud cluster.
- Huge advantage with Alces Flight.
  - Environment largely cloud ready
- Cloud prices vary with provider and time needed etc.
  - AWS better this test period.
  - Sacrificed faster interconnect for more nodes
- Plan – basic system login node, storage, small test node available several days before and after outage.

System configuration

- Login node – Skylake 24 cores, 350 GB memory
- 10 TB shared storage for all nodes
- 2 x 2C/16GB small compute nodes for testing
- 1 x Single Nvidia V100 GPU compute node
- 20 x 36C/128GB Skylake compute nodes
  - Only available just before poweroff and for 3 days after
- 100 GB of data / day down load from cluster
- Whilst local system up, easy to copy files.
- Used existing usernames with ssh keys for access.
- Home filestore **NOT** copied over.

Important considerations

- Our tailored Alces Flight Gridware preinstalled, we needed to copy over local application files.
  - Local module files completely replicated on cloud system
- Emphasis was on SMP parallel or coarse grain parallel plus Deep Learning on GPU.
- Nodes were hyperthreaded - users told to ask for exclusive access to avoid overloading.
- ssh key access – users told where to grab this from and the name of the cluster to ssh to.
- Cluster was only accessible from on-campus or via VPN to local system and then to cloud.

Lessons learned

- Once access obtained, people had no problem editing slurm scripts to run jobs.
- ssh keys, off-campus access slight niggles.
- Fewer people than expected used the system (only over a weekend).
  - Next time – check on how many likely users there will be.
  - Had more compute nodes than necessary; gpu node was used
- Cost of main compute nodes ~ 75% of overall cost
  - GPU node ~ 20%
- Similar look and feel to local system big help.
- Test nodes not helpful because lacked AVX-512

The major October outage

- Cloud cluster front end available from Friday 18/10
  - Easy connection / copy from on premise system
- Full cloud from 21st – powered off local compute
- Local storage / front-end power off 23rd.
- Local service restored on 30th October.
- Dropped cloud compute as jobs finished.
- Cloud front-end and storage kept available for several days so files could be transferred back.

Cloud cluster configuration - hardware

- Wanted 100 gbps InfiniBand
  - Got 100 gbps Ethernet
  - Performance was a bit erratic – not 100% certain why
- Started with 16 Cascade Lake (36 cores) with 4 AMD nodes for codes without AVX-512 builds plus v100 GPU.
  - AMD nodes not being used so went with 22 Cascade Lake nodes and just one AMD node.
- Could power off/on nodes to match demand
- 10 TB shared storage across cloud cluster

Cloud cluster configuration - environment

- Users connected using standard username and password (serviced by Active Directory on campus)
- Main login system appeared to be on Alces network
- Node and login images as per local system
- User home storage copied over in advance
- Module files largely worked as normal
- Tweaks to slurm scripts needed
- Had preliminary period for file-upload
- Appliance on campus used to channel AD requests

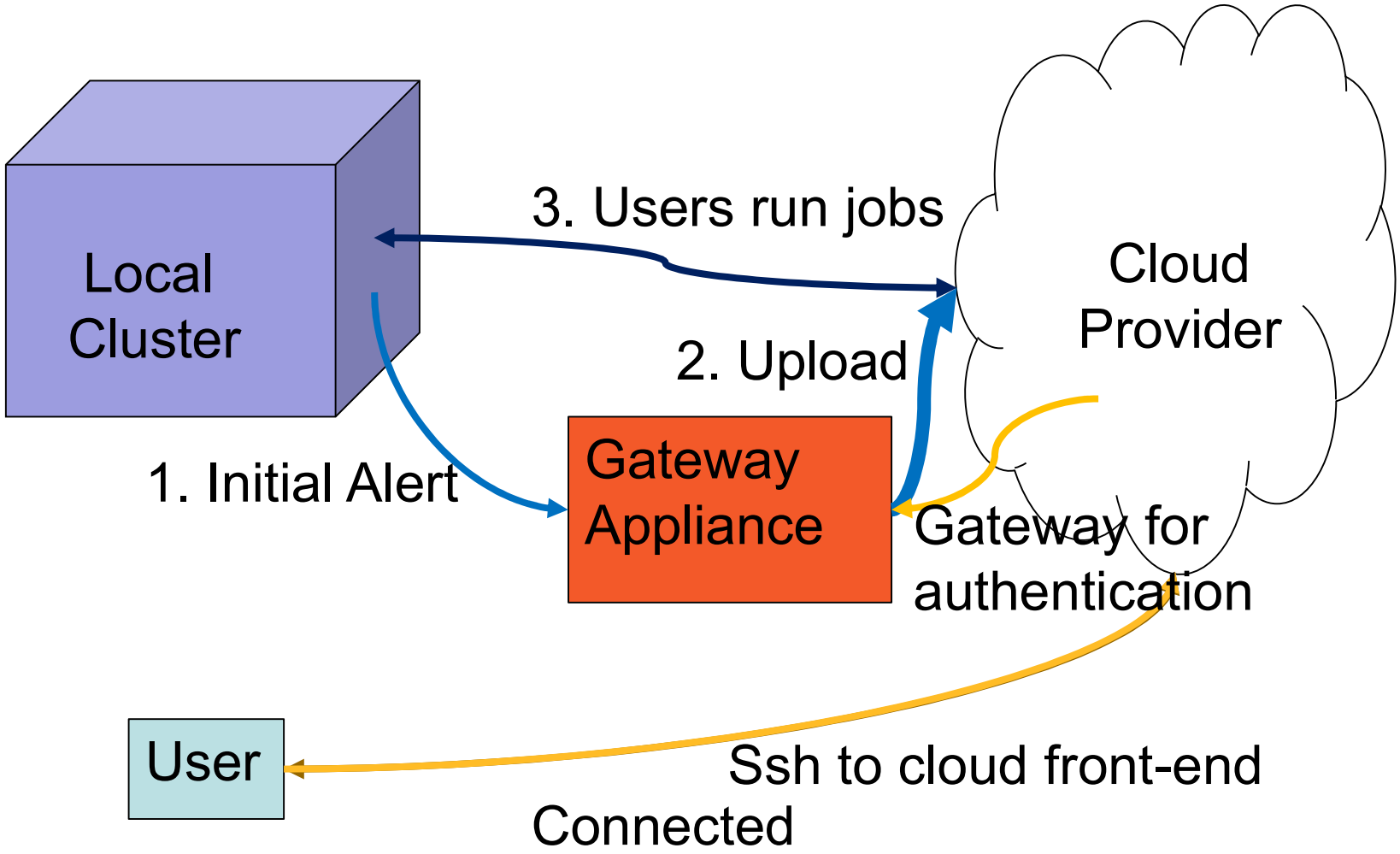
Use over the period

- Cascade Lake nodes very heavily used.
  - AMD node not used much at all
- GPU node constantly used after first couple of days
- Resources used
  - 278 user sessions to login node (19 individuals)
  - 275 GB new data generated
  - 560 slurm jobs processed (including 23 x GPU jobs)
  - 411GB data in / 275GB data out
- Mixture of SMP and small MPI jobs
- Cloud cost circa £20,000 (plus Alces logistical cost)
  - Roughly £2000 per day for 800 cores – fifth normal capacity

Plans after data centre move

- Local appliance functionality can be expanded.
  - Backup copy of node images, user data and orchestrate failover
- Integrate cloud cluster with university network
  - VPC so cluster appears as on the University network
- Bring up cloud cluster alongside Barkla to test “easy access” and cloud bursting potential.
- Experiment with spot market on AWS
  - Massive savings but small number of nodes
- Get firm University budget to sustain resiliency – local appliance plus occasional compute
  - Compute costs for full cluster mount very quickly!!

Some ideas on more general bursting



Gateway Appliance Roles

- This is potentially the special sauce for bursting!
- Node images & some user data synchronised here
- That data is uploaded and nodes instantiated on demand
  - No permanent cloud presence → no lock in!
- Users run jobs on cloud via on-premise cluster
- Also potential for direct connection to cloud cluster.
  - Gateway appliance provides authorisation and authentication
  - Can also be used to enforce other University policies.
- Potential to extend appliance to orchestrate other services (e.g. not just running HPC jobs)

Summary – general issues

- Need cloud cluster to have a similar look and feel to the local cluster.
- Integrate the cloud cluster into your local environment
  - Local appliance helps a lot
  - Active Directory / VPC so appears on campus network
- What storage is put where?
  - Local storage that is pushed to the cloud avoids lock-in – flexibility is good!
- Compute and login nodes created on-demand
  - How many compute nodes makes sense?? Interconnect??
  - Spot-market for some / all nodes
  - Ability to power on / off nodes is a must