

University of Birmingham & CLIMB GPFS User Experience

Simon Thompson

Research Support, IT Services

University of Birmingham



**UNIVERSITY OF
BIRMINGHAM**

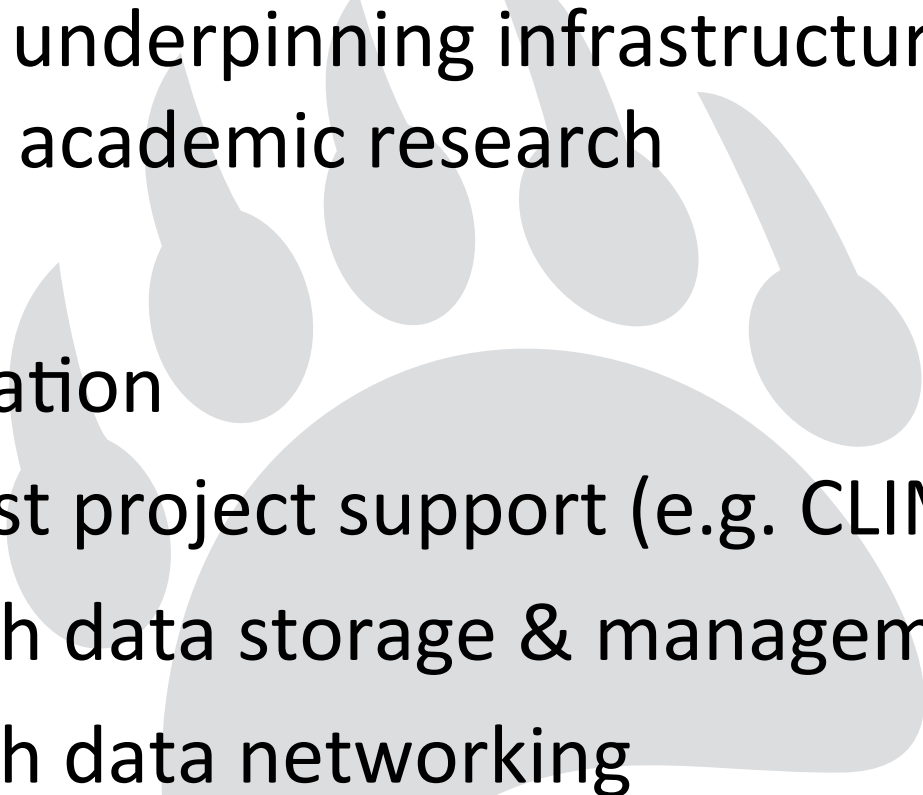
University of Birmingham

- Research intensive University
- ~19000 Undergraduate Students
- ~6400 Postgraduate Taught
- ~2900 Postgraduate Research
- £145.5 million in research income (2011-12)



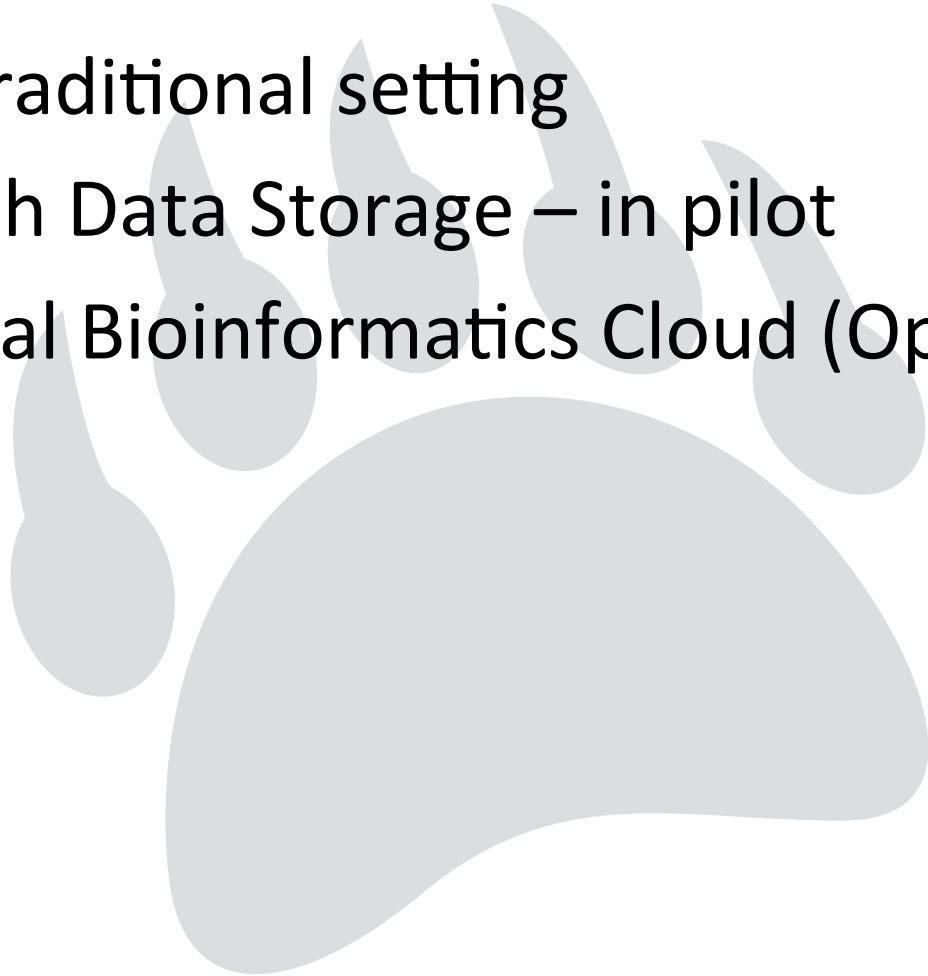
Data for 2011/2012 academic session

What does Research Support do?

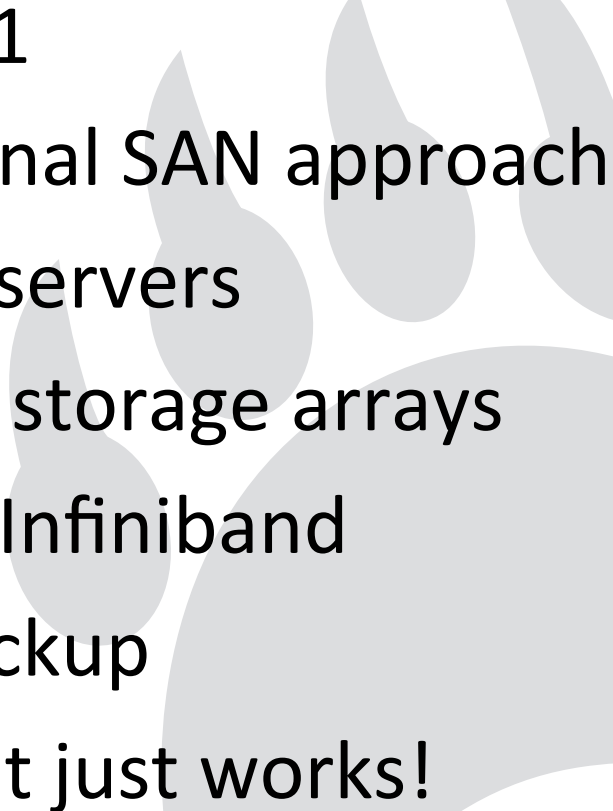
- Provide underpinning infrastructure to support academic research
 - HPC
 - Visualisation
 - Specialist project support (e.g. CLIMB)
 - Research data storage & management
 - Research data networking
- 

GPFS in action

- HPC – traditional setting
- Research Data Storage – in pilot
- Microbial Bioinformatics Cloud (OpenStack)

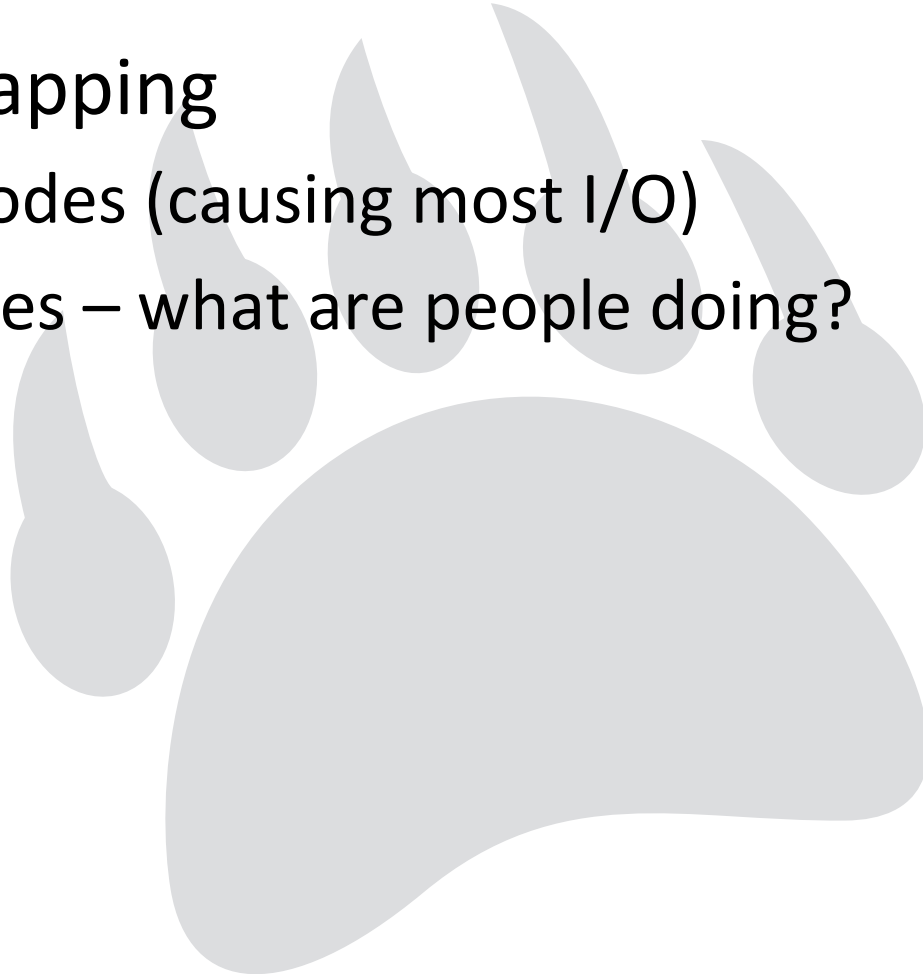


GPFS - HPC

- GPFS 4.1
 - Traditional SAN approach
 - 2x NSD servers
 - DS3500 storage arrays
 - FDR-10 Infiniband
 - TSM backup
 - Mostly it just works!
- 

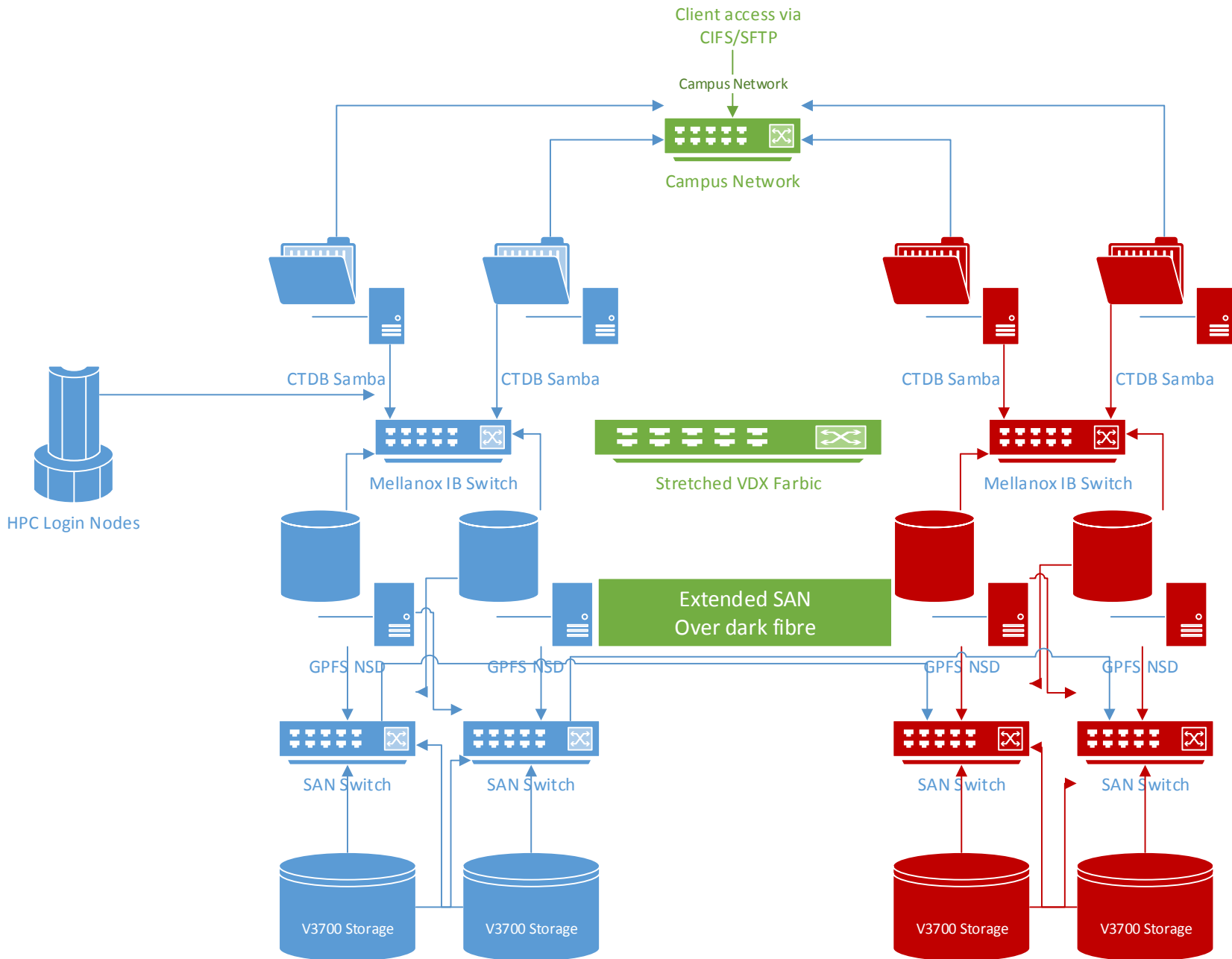
If I could change 1 thing in GPFS

- Heat mapping
 - Hot nodes (causing most I/O)
 - Hot files – what are people doing?



GPFS - Research Data Store

- Replicated across two data centres
- Separate IB fabrics at each DC
- 10GbE links between DCs
- Extended SAN based – users can buy space
- Designed and built as a partnership with OCF
- Engineered to be scale out



Research Data Store

- Clients access a separate Samba cluster
- Sernet Samba doesn't play with EL7 CTDB out of the box, but needed for GPFS VFS layer
- Had to patch the spec file and rebuild
- How will Samba play with HSM?
- Powerfolder sync and share pilot

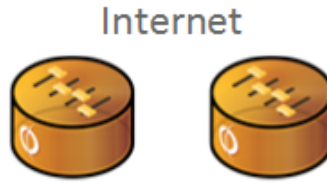
If I could change 2 things in GPFS

- Heat mapping
 - Hot nodes (causing most I/O)
 - Hot files – what are people doing?
- GNR should support scale out
 - Be able to add new shelves to a GNR

GPFS - OpenStack

- CLIMB is a Microbial Bioinformatics Cloud
- Funded by MRC, running on 4 sites (Birmingham, Warwick, Cardiff, Swansea)
- Allow users to archive VMs and datasets
 - Allows for re-use and sharing of data and code

vRouter provides encrypted ipsec tunnels between sites.



Brocade vRouter (HA Pair)

Mellanox Infiniband
GPFS traffic flows over IB using VERBS,
will fail back to Ethernet using subnets

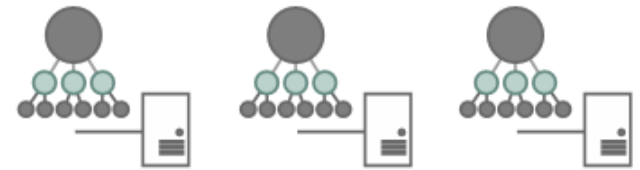


GPFS Servers (2-4)
Multiple v3700
storage arrays

Multiple fabric switches, all
hosts dual attached over 2
switches

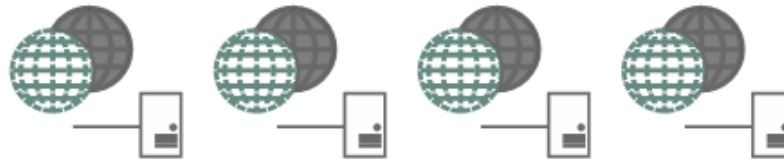


VLAN and VXLAN
networks



3 controller nodes
Mariadb cluster

haproxy, keepalived (VRRP) for services
native HA VRRP L3, multiple DHCP agents
for neutron



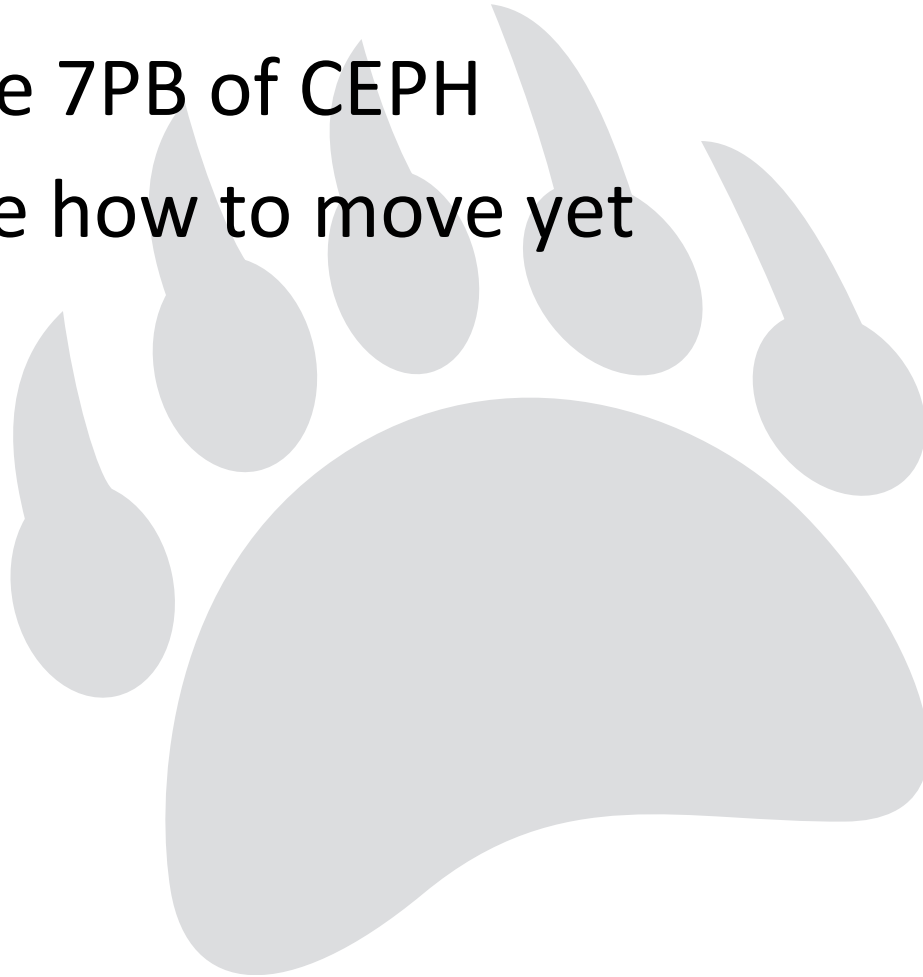
Compute / Hypervisor nodes
4 – 16 socket
512GB – 3TB RAM

How does GPFS fit in

- Swift (object store)
- Cinder/Glance driver interaction
 - Uses mmclone to rapidly provision VM images
- Nova ephemeral disks can reside on GPFS
 - Faster migration of VMs as shared storage

How do we archive VMs?

- We have 7PB of CEPH
- Not sure how to move yet



If I could change 3 things in GPFS

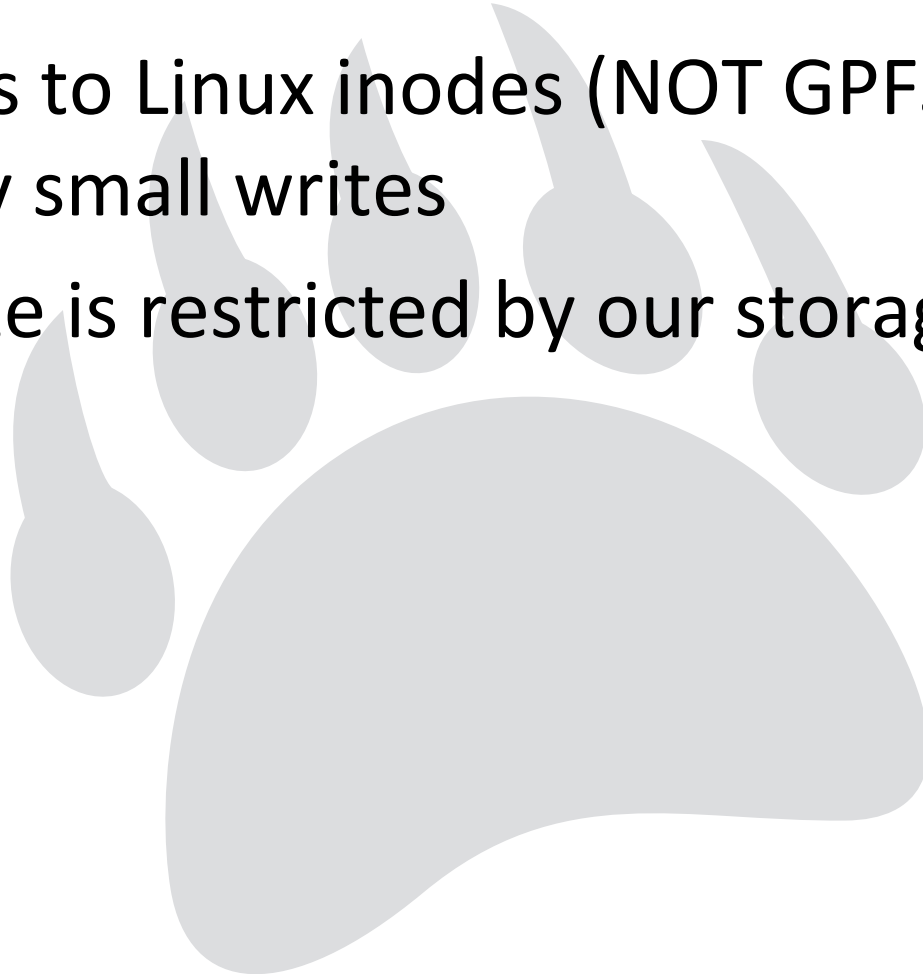
- Heat mapping
 - Hot nodes (causing most I/O)
 - Hot files – what are people doing?
- GNR should support scale out
 - Be able to add new shelves to a GNR
- **S3 HSM driver**
 - Our “stale” VMs would automatically migrate

OpenStack – metrics (Cinder driver)

- Time to from clicking launch to running state
- Boot from image - 1m 17 seconds
- Boot from image (new volume) - 19 seconds
- (CentOS 7 image, 8GB, RAM)

Small IO is a killer

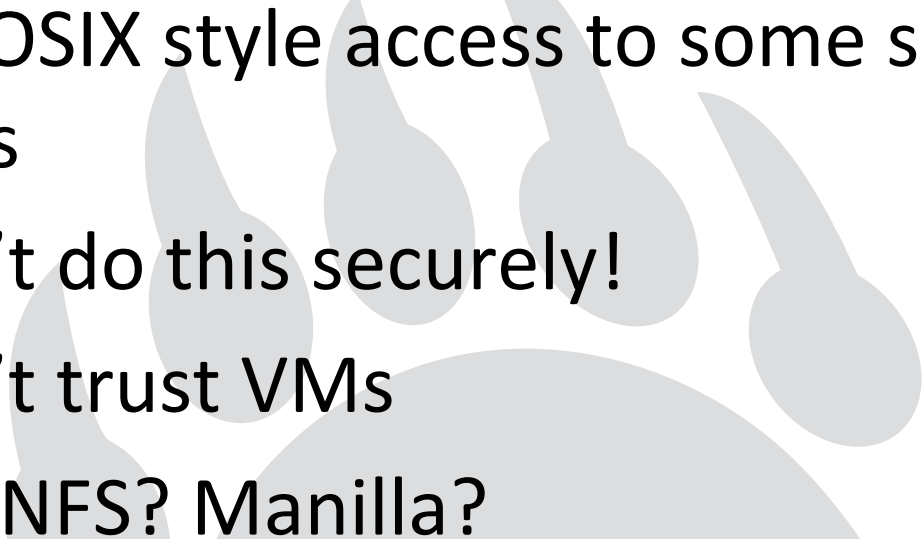
- Updates to Linux inodes (NOT GPFS inodes) are very small writes
- Blocksize is restricted by our storage arrays



If I could change 4 things in GPFS

- Heat mapping
 - Hot nodes (causing most I/O)
 - Hot files – what are people doing?
- GNR should support scale out
 - Be able to add new shelves to a GNR
- S3 HSM driver
 - Our “stale” VMs would automatically migrate
- Support a write caching layer for small IO
 - E.g. onto SSD pool

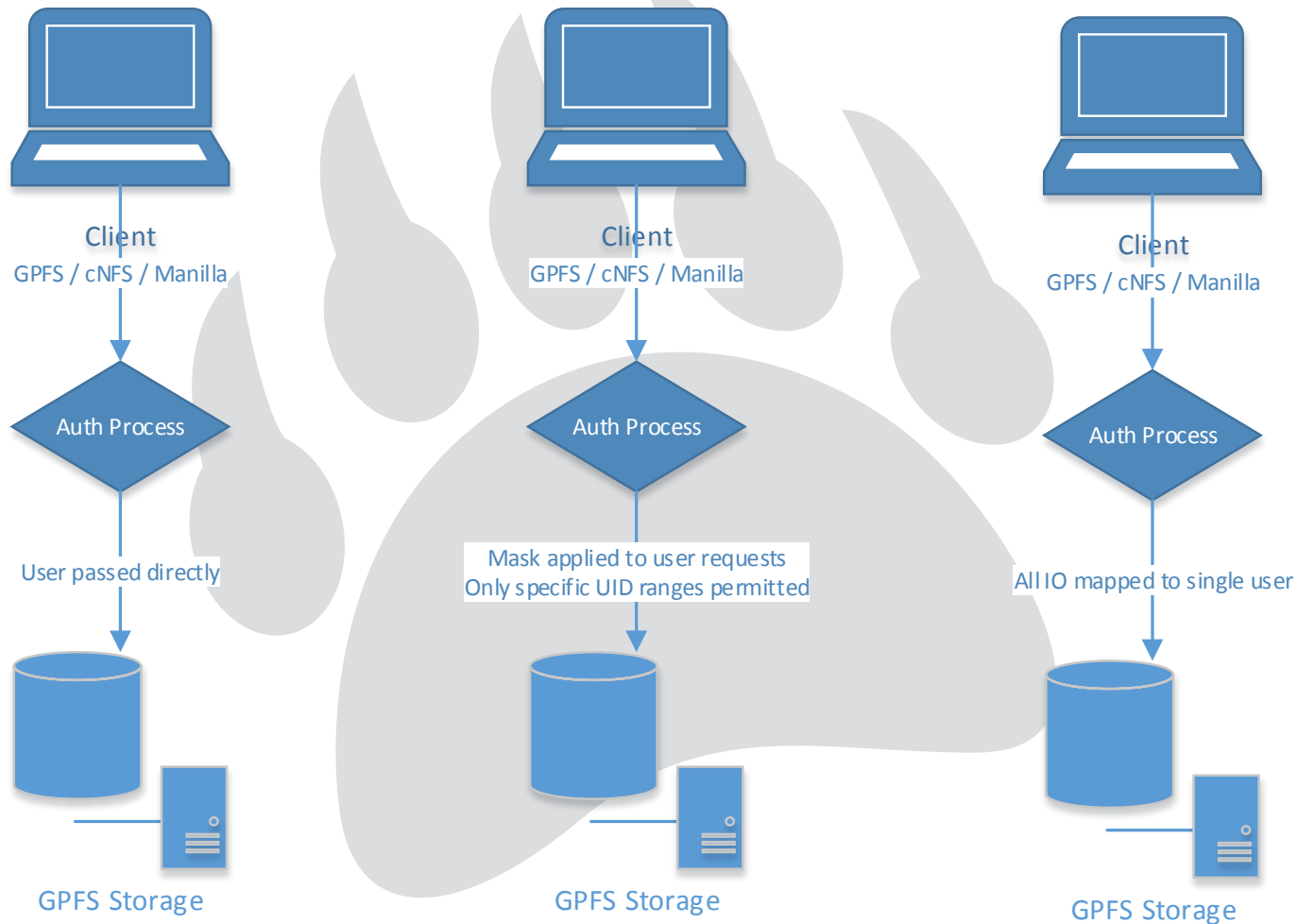
GPFS inside the VMs?

- Want POSIX style access to some shared datasets
 - We can't do this securely!
 - We can't trust VMs
 - GPFS? cNFS? Manilla?
- 

If I could change 5 things in GPFS

- Heat mapping
 - Hot nodes (causing most I/O)
 - Hot files – what are people doing?
- GNR should support scale out
 - Be able to add new shelves to a GNR
- S3 HSM driver
 - Our “stale” VMs would automatically migrate
- Support a write caching layer for small IO
 - E.g. onto SSD pool
- **Finer grained security for client access**

What might that look like?



Finally ...

- www.bear.bham.ac.uk
- www.climb.ac.uk
- www.roamingzebra.co.uk (my blog)



