BIOTEAM
Enabling Science

Converged&
Fault Tolerant&
Distributed&
Parallel&
iRODS.

iRODS User Group Meeting 2017

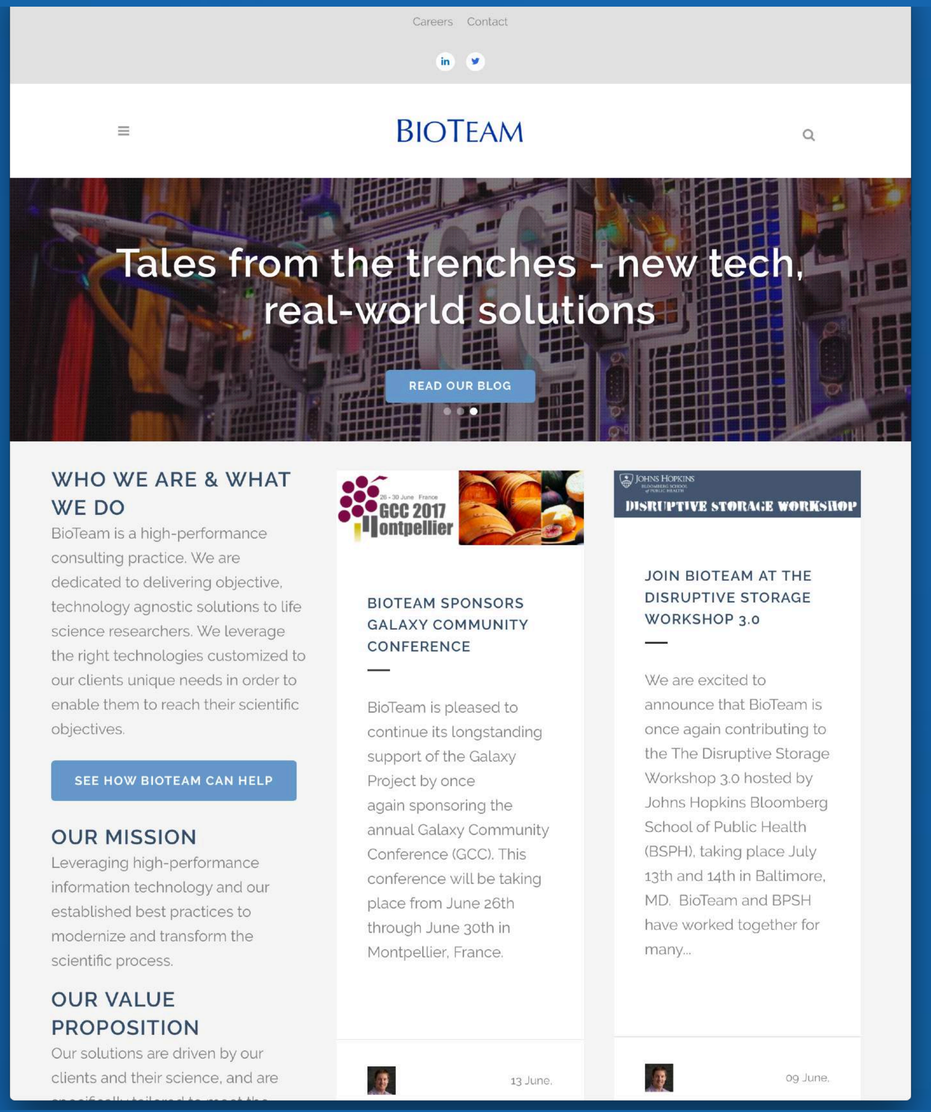
Aaron Gardner

June, 2017

Public - Distribution Permitted

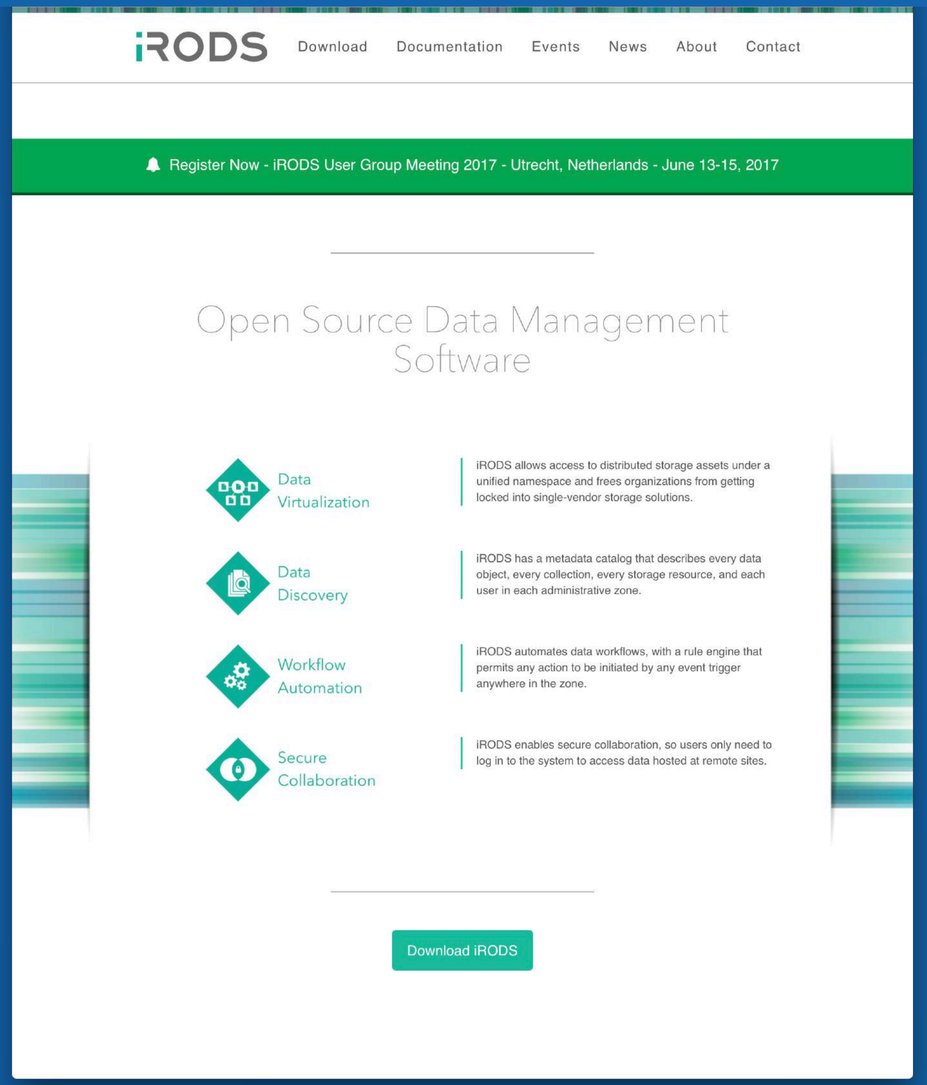
Introduction

- BioTeam is focused on research computing consulting and products
- Scientists with deep IT and scientific computing expertise
- Infrastructure (HPC, Storage, Networking, Enterprise, Cloud), Informatics, Software Development, Cross-disciplinary Assessments
- 15 years bridging the “gap” between science, IT, and HPC



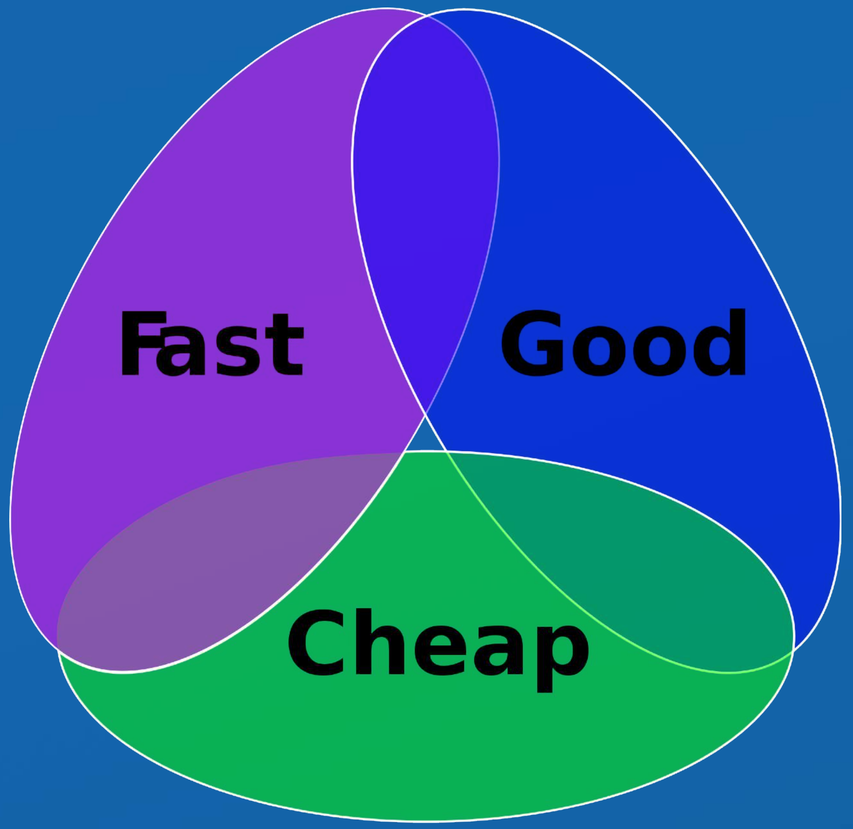
History with iRODS

- BioTeam members working with iRODS since 2011— thanks Reagan
- A number of consulting engagements around iRODS
- BioTeam sees data management as a critical mountain that must be scaled
- We are actively engaged with the scientific community to solve data management issues collaboratively



Motivation

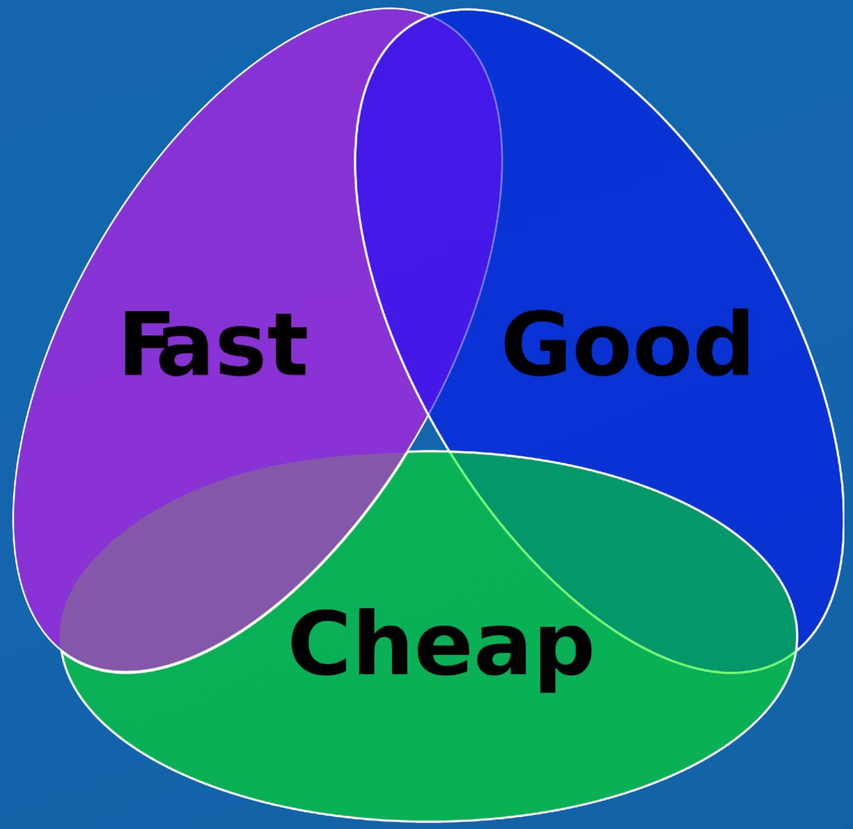
- Resource server vault storage exclusivity
- OK for direct attached storage, active archive
- Not for distributed parallel at speed
- Multiple copies on primary (fast) storage for iRODS a non-starter



Motivation

- Resource server fails—data drops off the grid
- Catalog fails—lose access to everything
- Multiple copies of catalog data not ideal
- Avoid additional hardware
- Performance and scalability

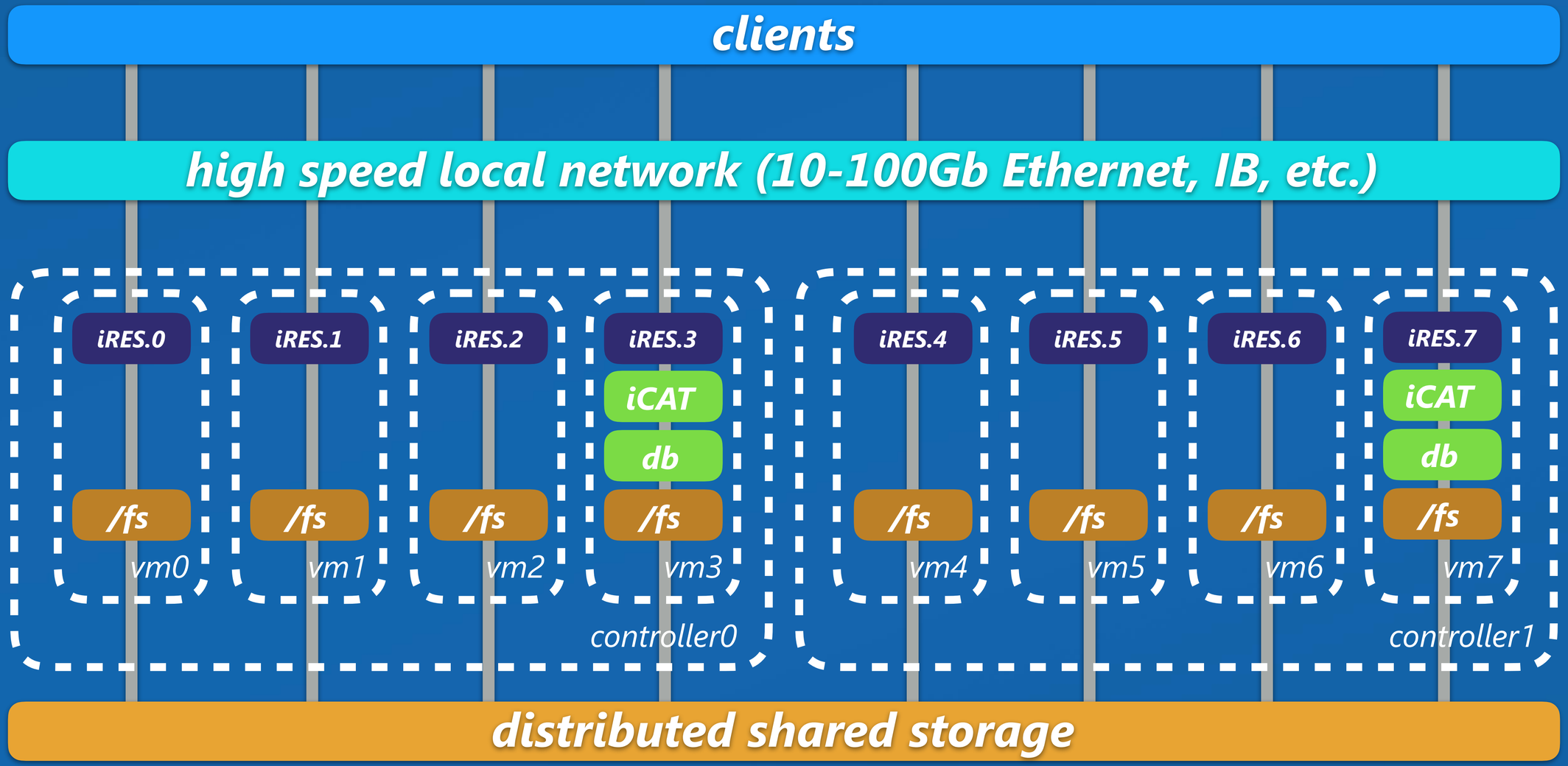
We want "all the things"—what to do?



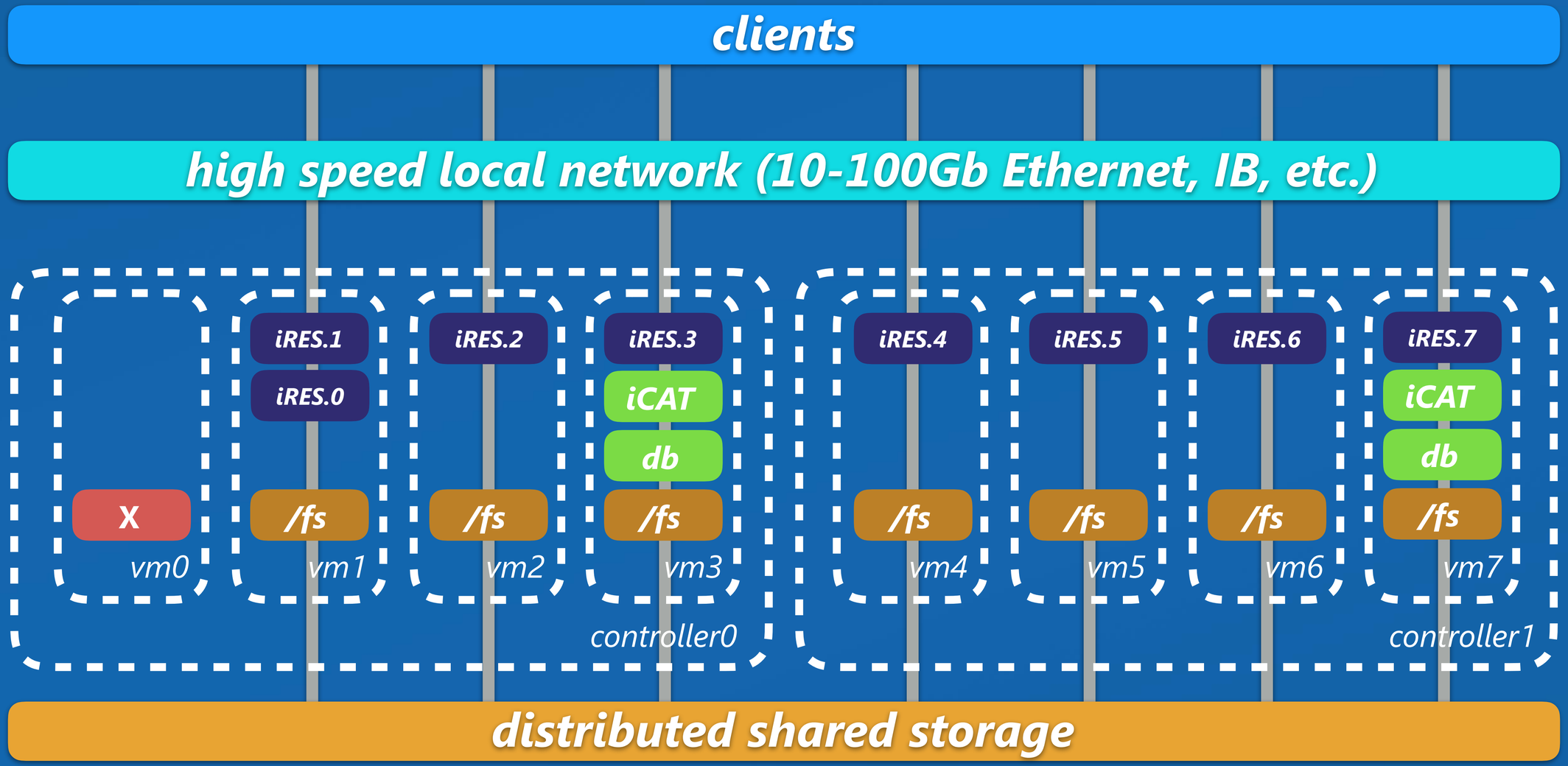
Can an iRODS catalog and resources have the same resiliency and scalability that today's distributed storage systems have?

How close can we get?

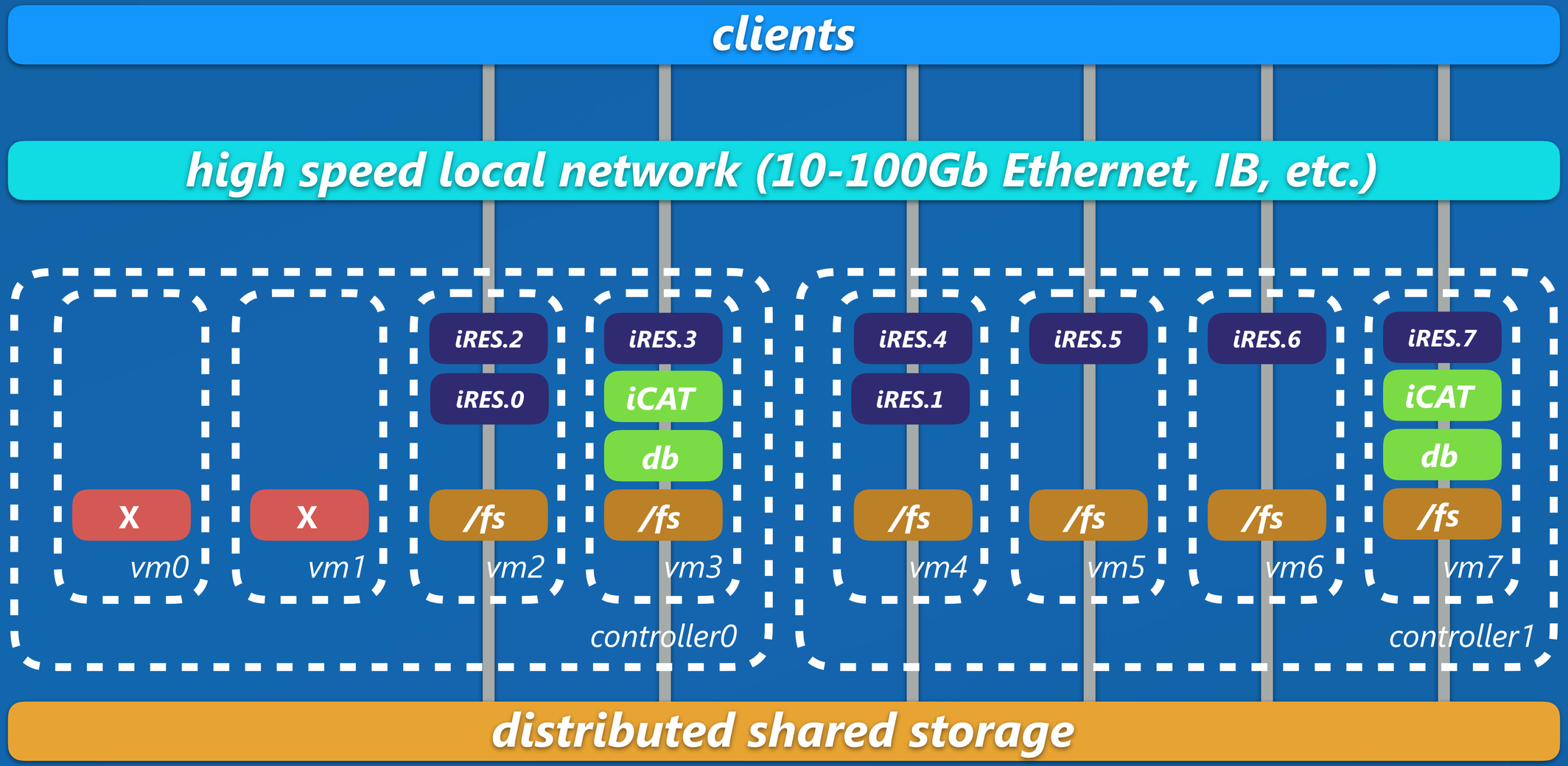
New Reference Architecture



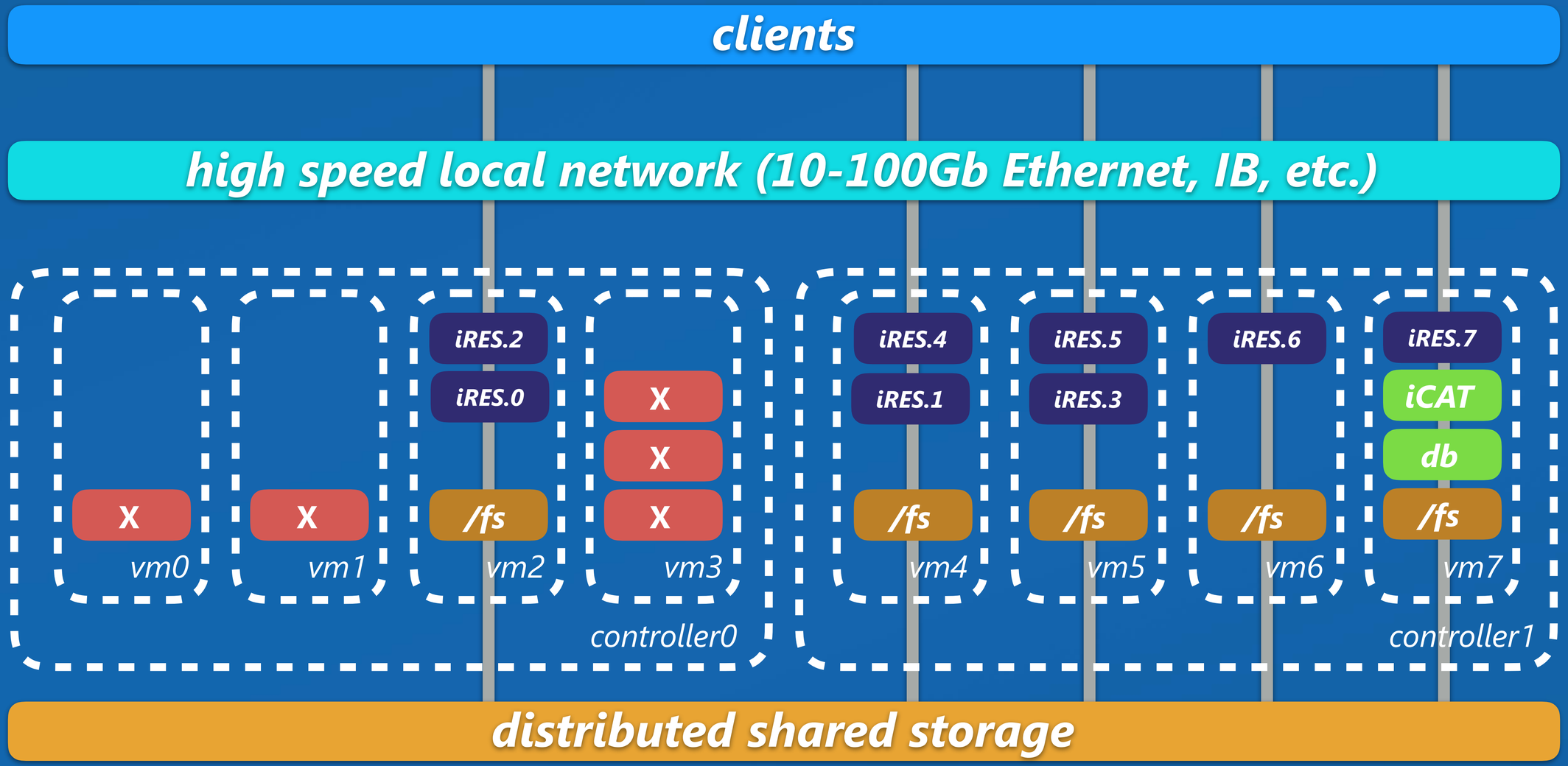
Let it fail



Let it fail, let it fail



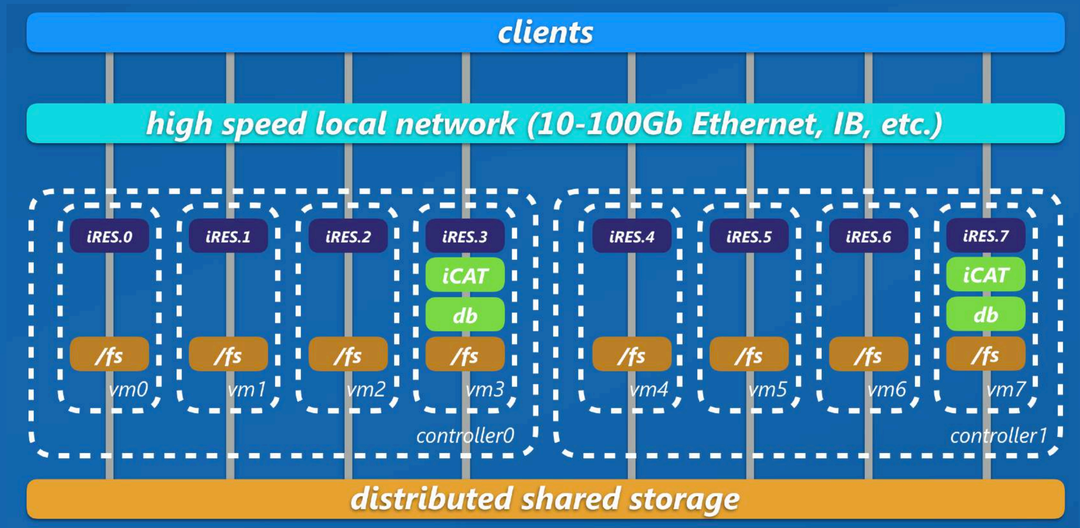
Let it fail, let it fail, let it fail.



New Reference Architecture

Converged

- Deployed on storage controller(s)
- No additional hardware or server instances
- Request latency minimized
- Single replica kept on shared storage



Fault Tolerant

- Resource servers see all available storage
- “Physical” resources impersonate “virtual”
- Cluster monitoring and failure handling
- Only need one “physical” resource, catalog, database

New Reference Architecture

Distributed

- Resource performance scales with backing storage
- iCAT hosted on distributed storage and scales independently

Parallel

- Client can read and write to all resources at the same time
- Minimize false “data island” lock-in
- Clients can achieve higher bandwidth than a single resource
- (Future) Multipart could provide true parallel object access

iRODS

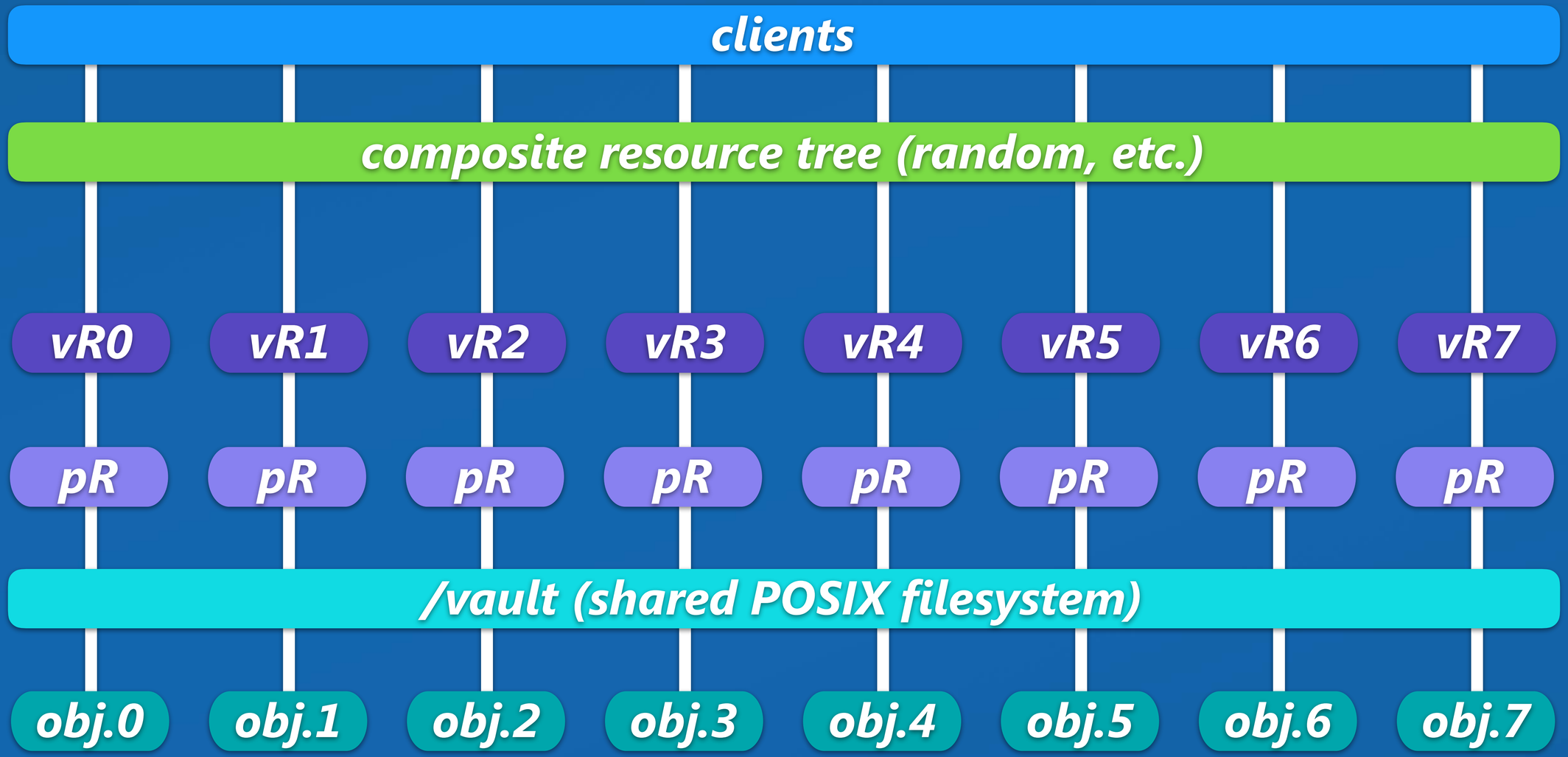
- Unmodified codebase
- Scale horizontally
- Incorporate with other storage

How was this accomplished?

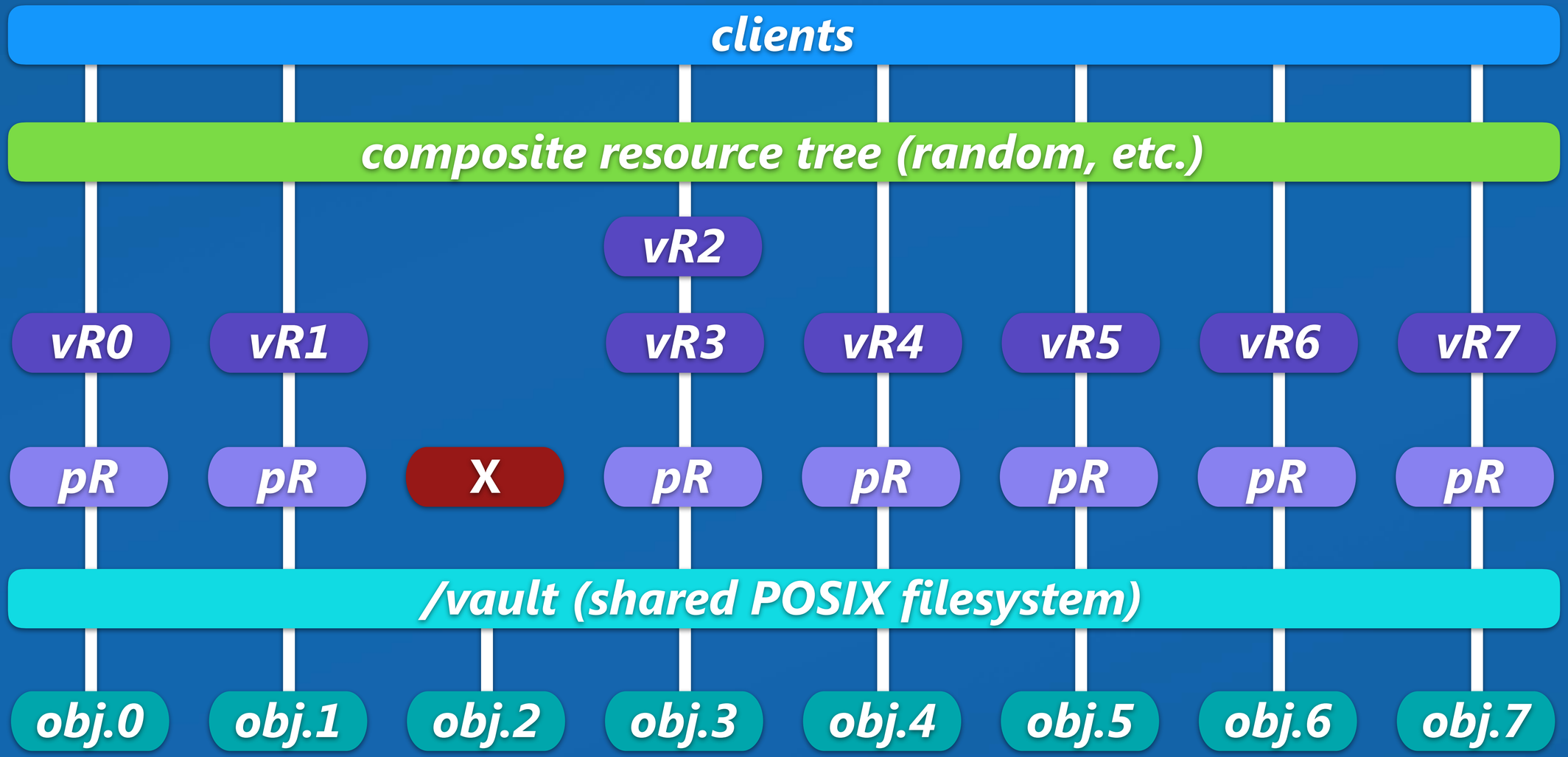
- iRODS 4.1.9 (refactoring for 4.2.1)
- Ansible, Vagrant, VirtualBox, NFS for Test
- Spectrum Scale on Cluster for Production
- Pacemaker/(CMAN | Corosync)
- Custom *irods*, *icat* OCF resources
- "Virtual" resource reference counting
- */etc/irods/hosts_config.json*
- Galera Cluster for MySQL

```
15  - name: pacemaker | 0001 | Generate {{ hacluster_config.file }}
16    template: src=templates/pacemaker/cluster.conf.j2 dest={{ hacluster_config.file }}
17    when: host_is_resource_server == True or host_is_icat_server == True
18
19  - name: pacemaker | 0002 | Create {{ hacluster_irods_pacemaker_resource }} Directory
20    file: path={{ hacluster_irods_pacemaker_resource_dir }} state=directory
21    when: host_is_resource_server == True or host_is_icat_server == True
22
23  - name: pacemaker | 0003 | Copy iRODS Pacemaker Resource Script
24    template: src=templates/pacemaker/irods.j2 dest={{ hacluster_irods_pacemaker_resource }} mode=0755
25    when: host_is_resource_server == True or host_is_icat_server == True
26
27  - name: pacemaker | 0004 | Change Quorum Timeout
28    lineinfile: dest=/etc/sysconfig/cman line="CMAN_QUORUM_TIMEOUT=0" insertafter=EOF
29    when: host_is_resource_server == True or host_is_icat_server == True
30
31  - name: pacemaker | 0005 | Enable Pacemaker Services
32    service: name={{ item }} state=started enabled=yes
33    with_items:
34      - pcsd
35      - cman
36      - pacemaker
37    when: host_is_resource_server == True or host_is_icat_server == True
38
39  - name: pacemaker | 0006 | Change {{ hacluster_user }} Password
40    user: name={{ hacluster_user }} password={{ hacluster_user_password_crypted }} update_password=always
41    when: host_is_resource_server == True or host_is_icat_server == True
42
43  - name: pacemaker | 0007 | Update pcs Cluster Auth
44    command: pcs cluster auth -u {{ hacluster_user }} -p {{ hacluster_user_password }} {{ item.name }}
45    with_items: "{{ hacluster_config.nodes }}"
46    when: host_is_resource_server == True or host_is_icat_server == True
47    register: pcs_cluster_auth
48    failed_when: pcs_cluster_auth.rc > 1
49    changed_when: False
```

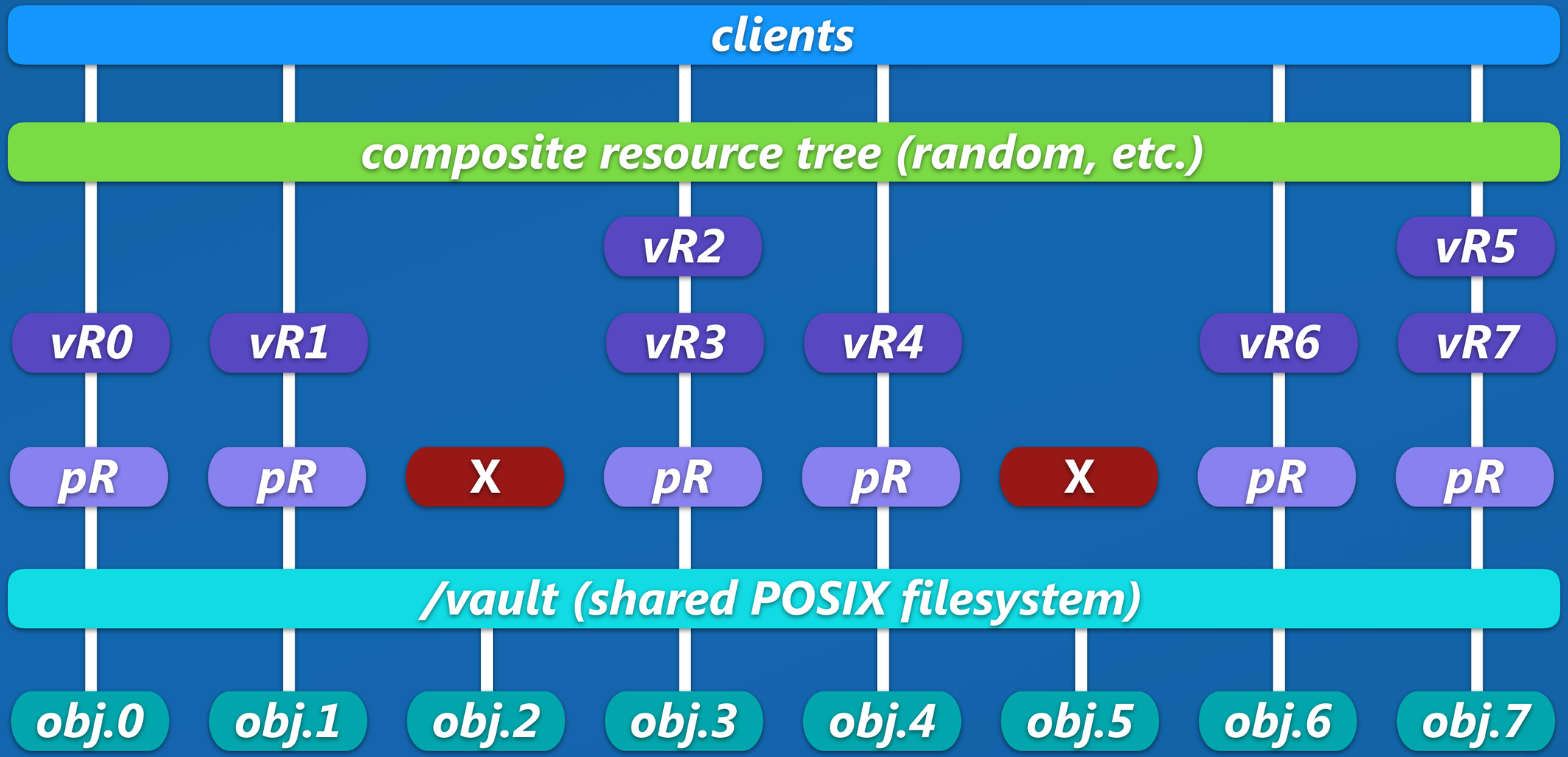
Physical Resource *(pR)* Failures: *0*



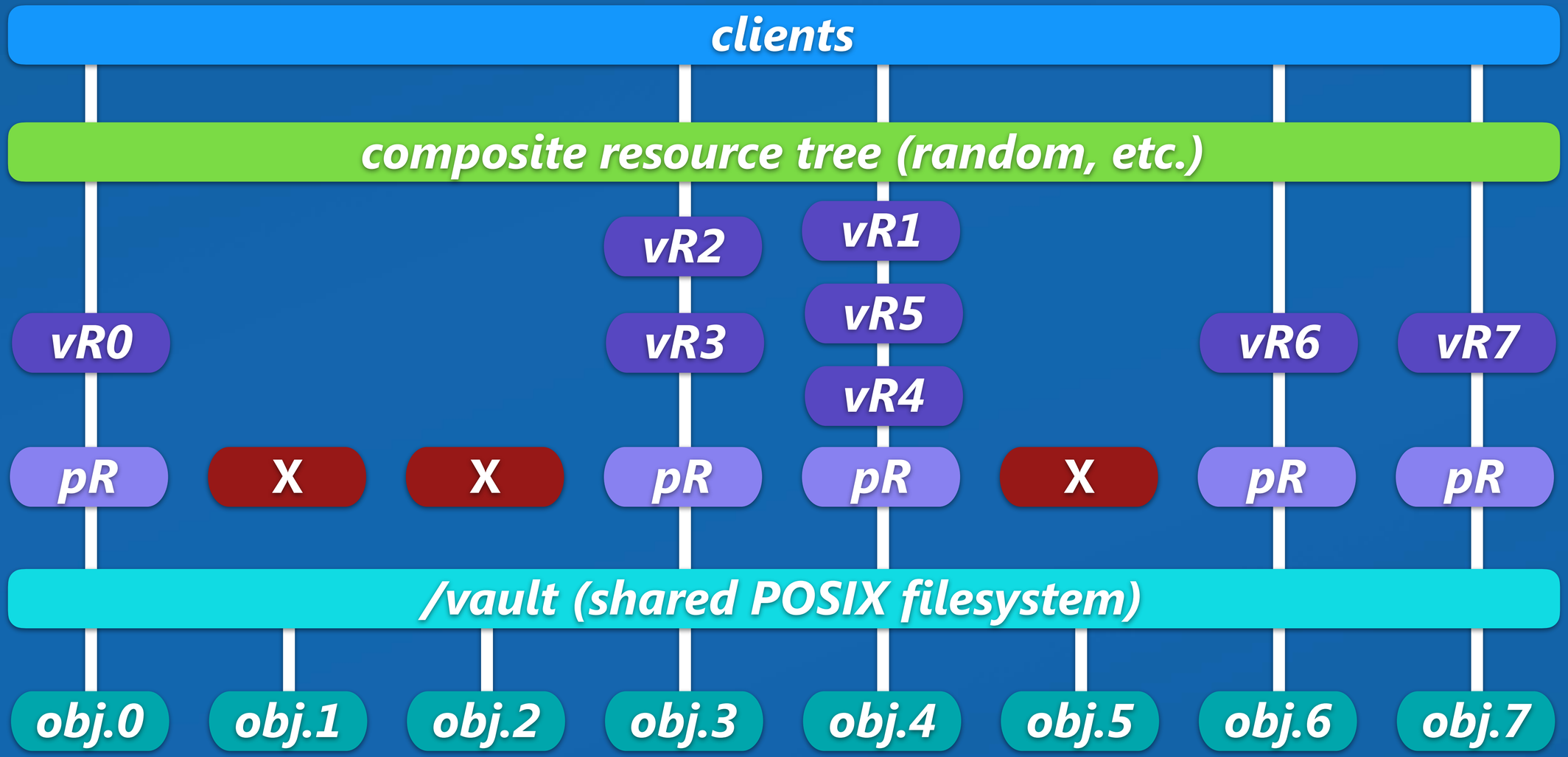
Physical Resource *(pR)* Failures: *1*

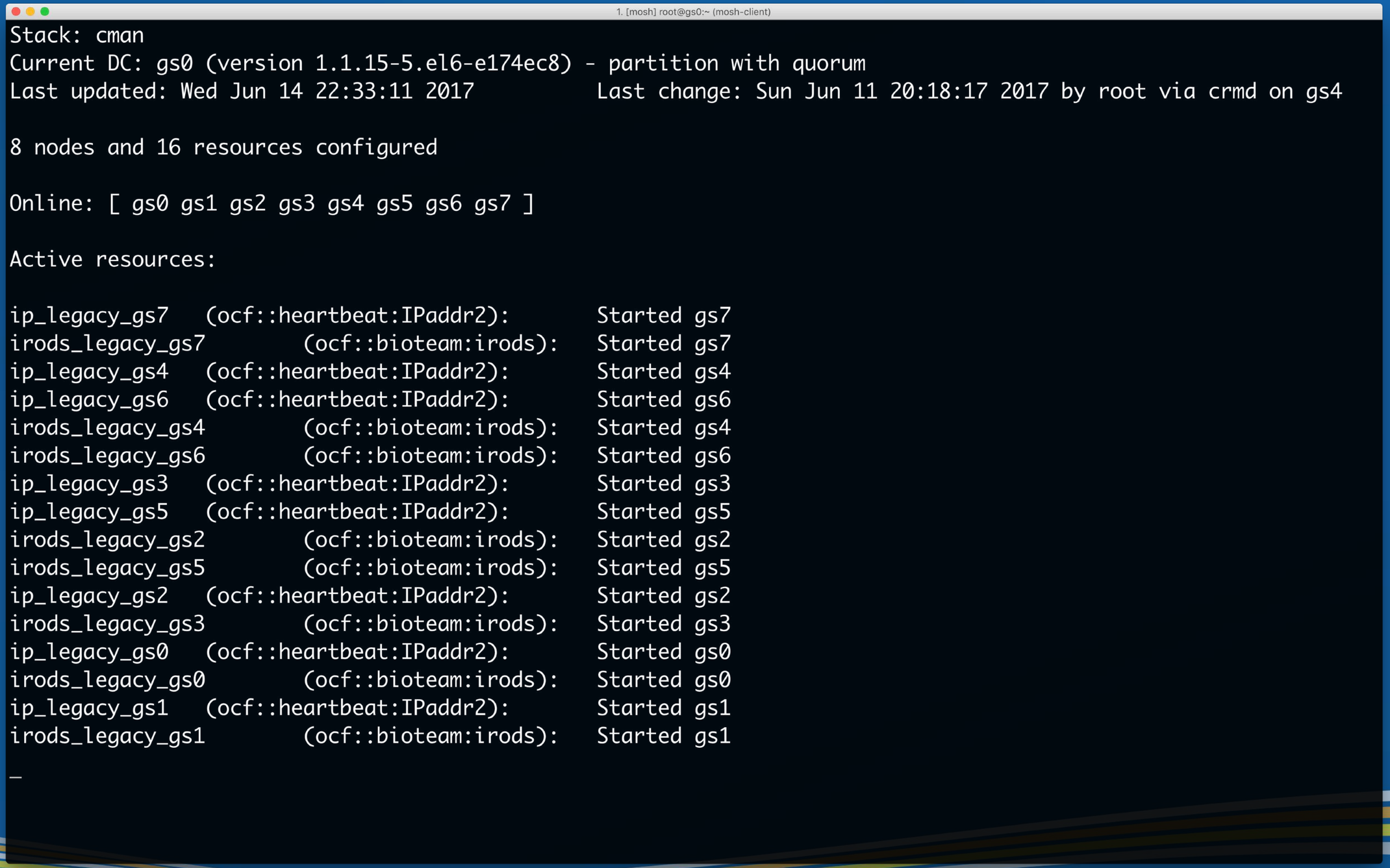


Physical Resource *(pR)* Failures: *2*



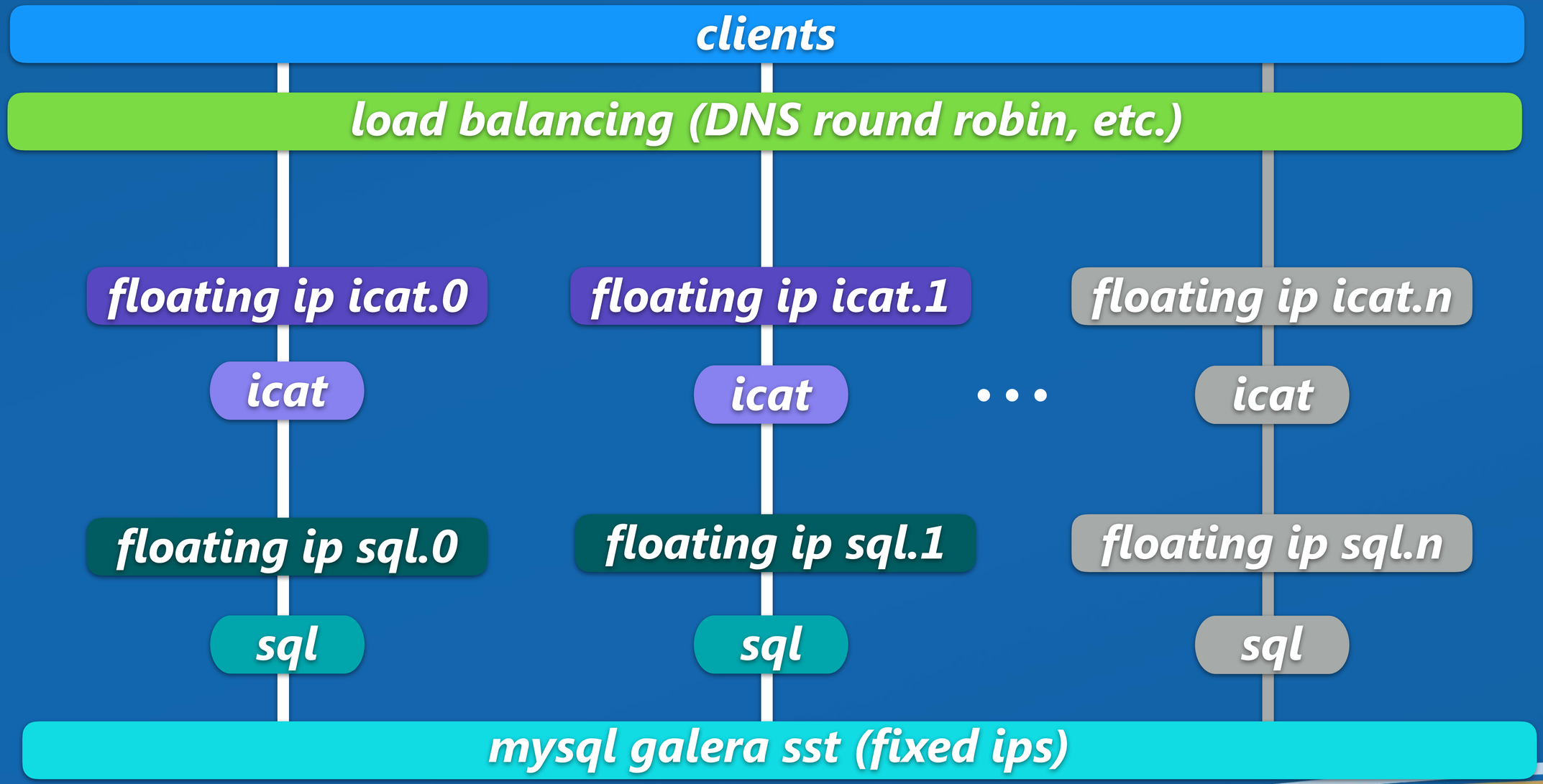
Physical Resource *(pR)* Failures: *3*



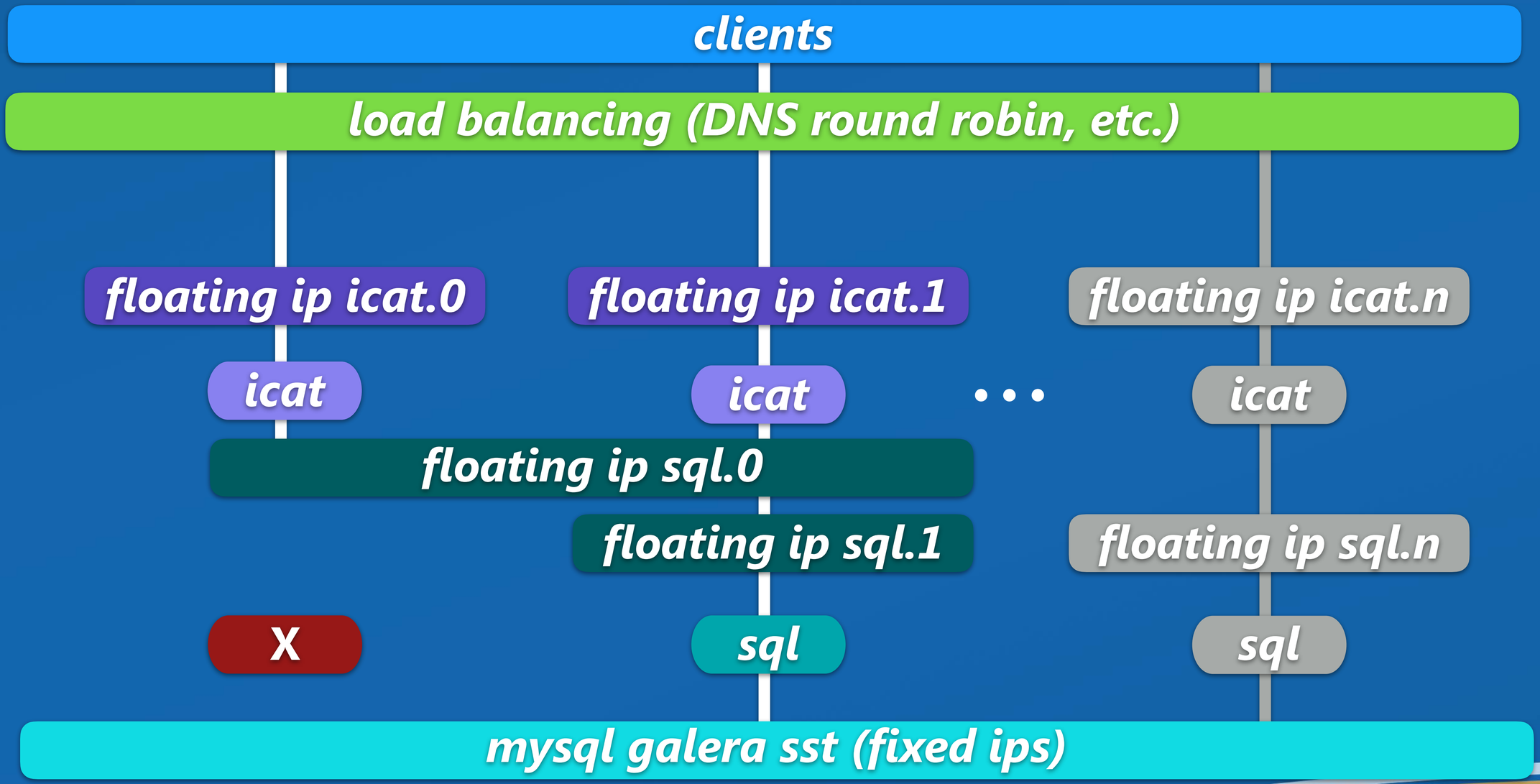


```
Stack: cman
Current DC: gs3 (version 1.1.15-5.el6-e174ec8) - partition WITHOUT quorum
Last updated: Wed Jun 14 22:40:03 2017          Last change: Sun Jun 11 20:18:17 2017 by root via crmd on gs4

8 nodes and 16 resources configured

Online: [ gs3 ]
OFFLINE: [ gs0 gs1 gs2 gs4 gs5 gs6 gs7 ]

Active resources:

ip_legacy_gs7   (ocf::heartbeat:IPaddr2):       Started gs3
irods_legacy_gs7        (ocf::bioteam:irods):   Started gs3
ip_legacy_gs4   (ocf::heartbeat:IPaddr2):       Started gs3
ip_legacy_gs6   (ocf::heartbeat:IPaddr2):       Started gs3
irods_legacy_gs4        (ocf::bioteam:irods):   Started gs3
irods_legacy_gs6        (ocf::bioteam:irods):   Started gs3
ip_legacy_gs3   (ocf::heartbeat:IPaddr2):       Started gs3
ip_legacy_gs5   (ocf::heartbeat:IPaddr2):       Started gs3
irods_legacy_gs2        (ocf::bioteam:irods):   Started gs3
irods_legacy_gs5        (ocf::bioteam:irods):   Started gs3
ip_legacy_gs2   (ocf::heartbeat:IPaddr2):       Started gs3
irods_legacy_gs3        (ocf::bioteam:irods):   Started gs3
ip_legacy_gs0   (ocf::heartbeat:IPaddr2):       Started gs3
irods_legacy_gs0        (ocf::bioteam:irods):   Started gs3
ip_legacy_gs1   (ocf::heartbeat:IPaddr2):       Started gs3
irods_legacy_gs1        (ocf::bioteam:irods):   Started gs3

_
```

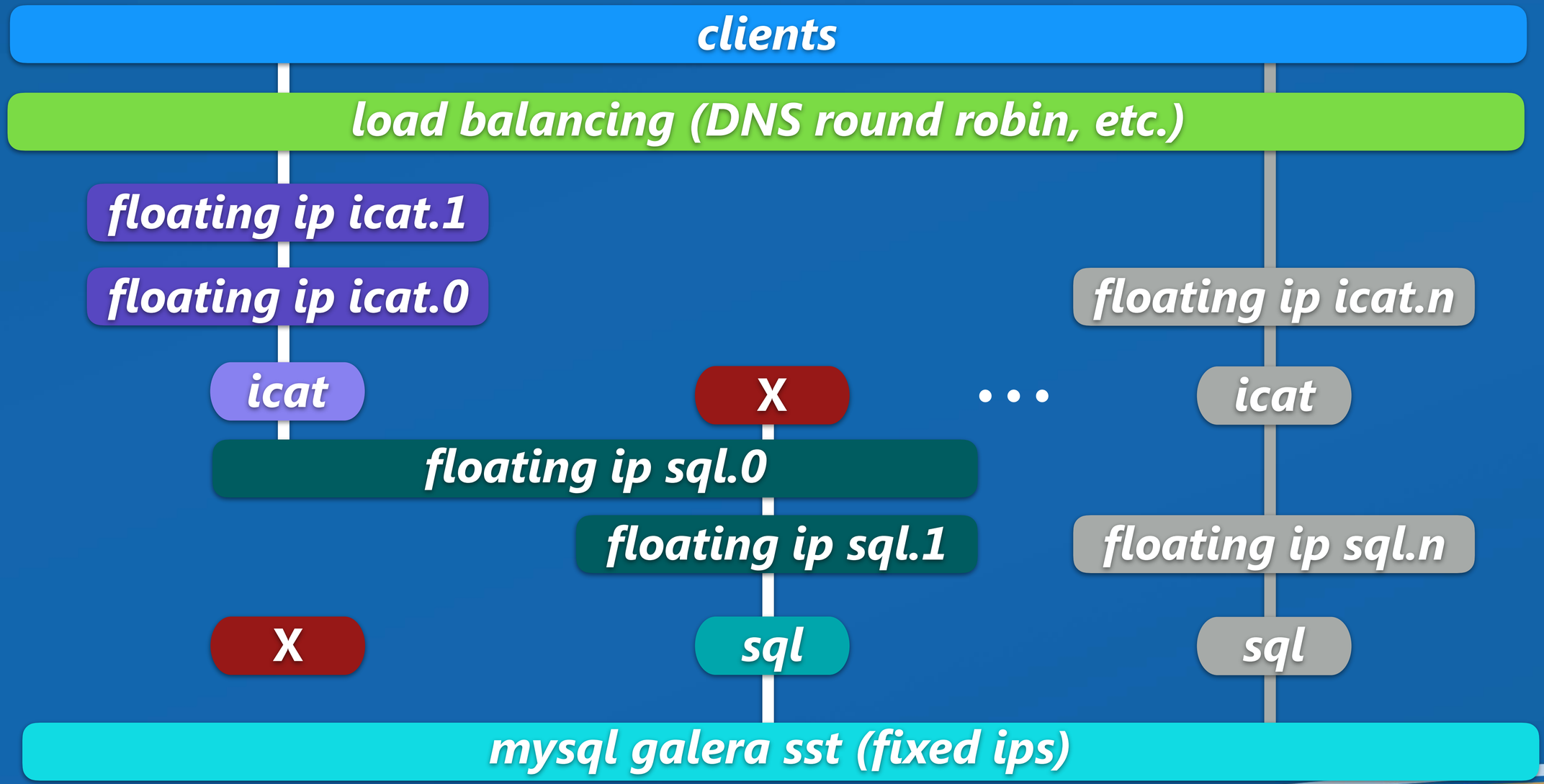
HA Active-Active iCAT Cluster



HA Active-Active iCAT Cluster: *SQL Fail*



HA Active-Active iCAT Cluster: *iCAT Fail*



BIOTEAM
Enabling Science

iRODS Distributed Database Experiences

Oracle RAC

MySQL Cluster

Postgres-XL

MySQL Galera

iRODS Soapbox

- Resource throughput and scalability
- Catalog performance and scalability
- Atomicity of transactions
- Multipart
- Multipath for resources
- Fastpath

BIOTEAM
Enabling Science

Future Work

- Benchmark and test
- Postgres-XL
- Apache Trafodion
- Desirable replication
- Additional architectures (HCI, etc.)
- Microservice deployment in Kubernetes

Thank You

bioteam.net
info@BioTeam.net
@BioTeam