HIGHAVAILABILITY

OpenZFS

Paris 2014

Grenville Whelan

High-Availability.com

High-Availability.com

- Spun-off from Sun/Solaris Consulting Practice in 1995
- RSF-1 first independent commercial HA solution for UNIX
- Heritage in enterprise SPARC/Solaris/Oracle and Linux
- 19 year proven track record and constantly evolving
- “Keep It Simple” design ethos
- Global install base includes: Emergency Services, Government, Banking, Stock Exchanges, Retail, Engineering, Telco
- Thousands of enterprise deployments around the world

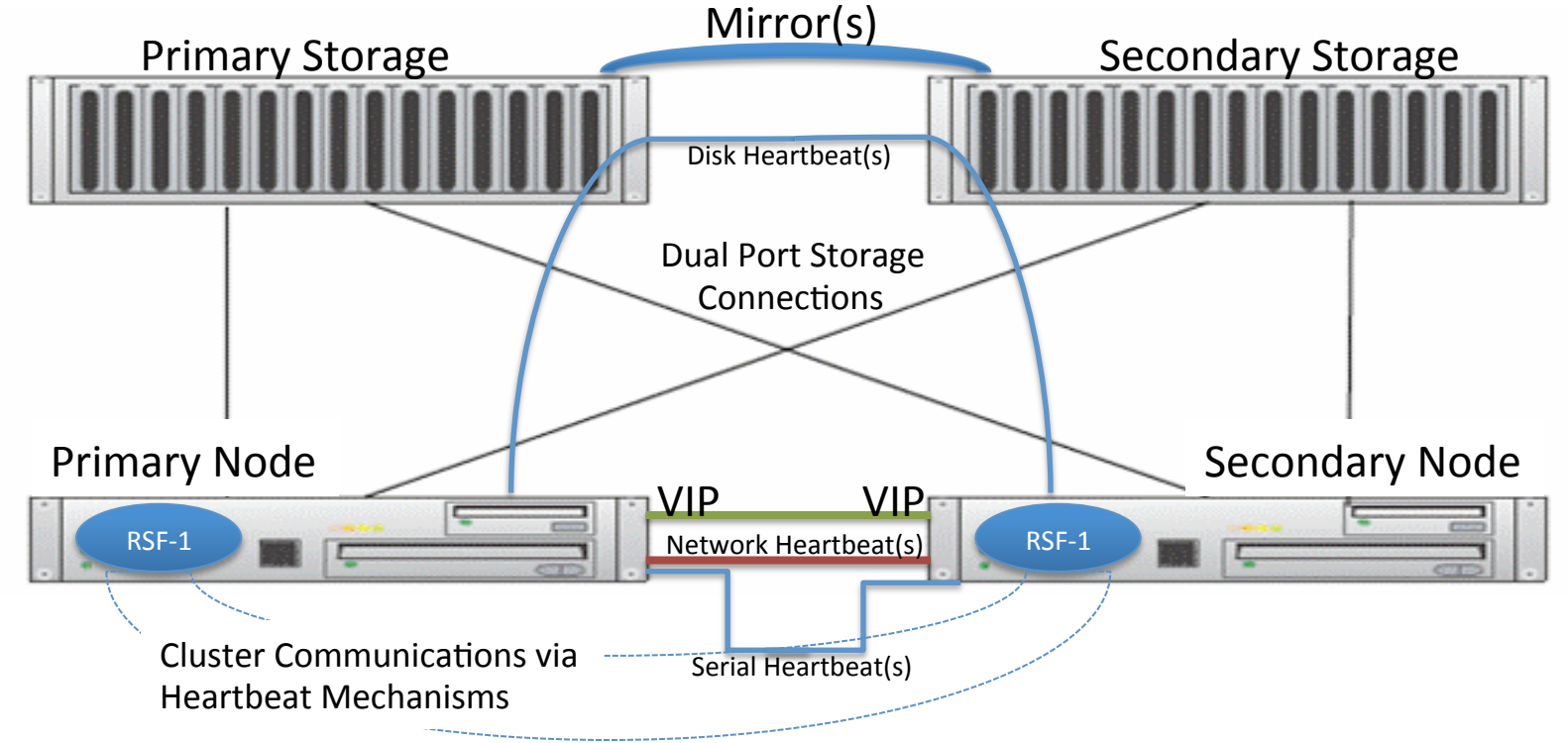
RSF-1 HA Capabilities

- Supports 1-256 services and 2-64 node clusters
- Platform support: Solaris, Open Indiana, illumos, OmniOS, SmartOS, FreeBSD, Linux (and others)
- Application framework for third-party hardware, applications, middleware and databases
- Flexible and easy to configure – Active/Active/N, Active/Passive
- Advanced “stretch cluster” capabilities for HA/DR
- Automatic/Manual failover of individual services
- Standalone GUI & Command Line Interface
- Cluster-wide state information propagated via heartbeats
- Dynamic single-point cluster-wide configuration with hot restart
- Very light-weight process(es), minimal system overhead

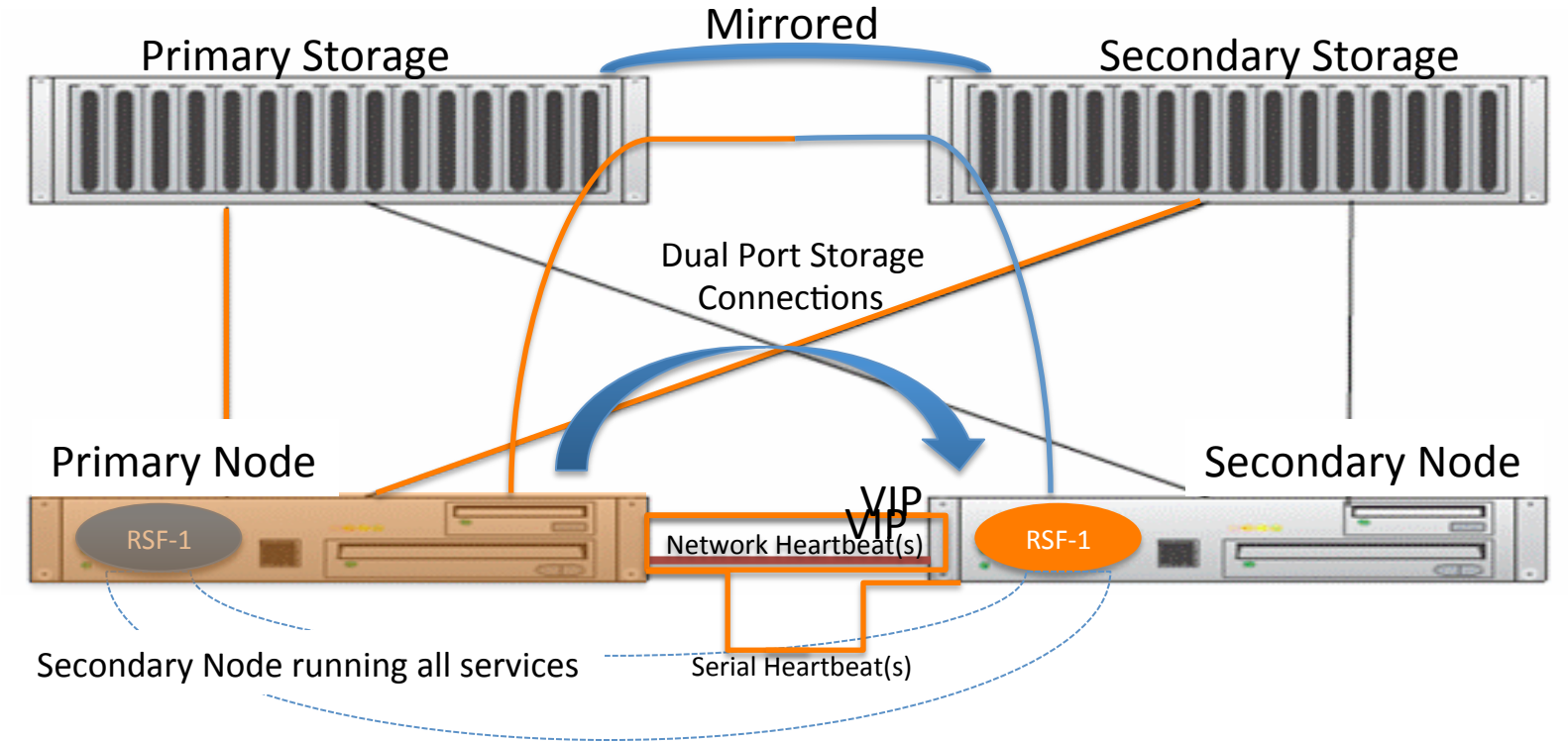
OpenZFS “HA Plugin”

- HA Cluster plugin for many ZFS vendors (Nexenta, Coraid ..)
- More than 2,000 ZFS based clusters deployed to date
- Seamless failover of Block/File services (COMSTAR,ALUA,NFS, CIFS)
- Built-in disk-fencing protection (SCSI-2/PGR3) mechanisms
- Fault Management Integration failover capability
- Optimised ZFS (import/export) failover time
- System freeze detection and protection

RSF-1 Typical 2-node Topology



Primary Node Failure



RSF-1 for OpenZFS

As OpenZFS HA Plugin, RSF-1:

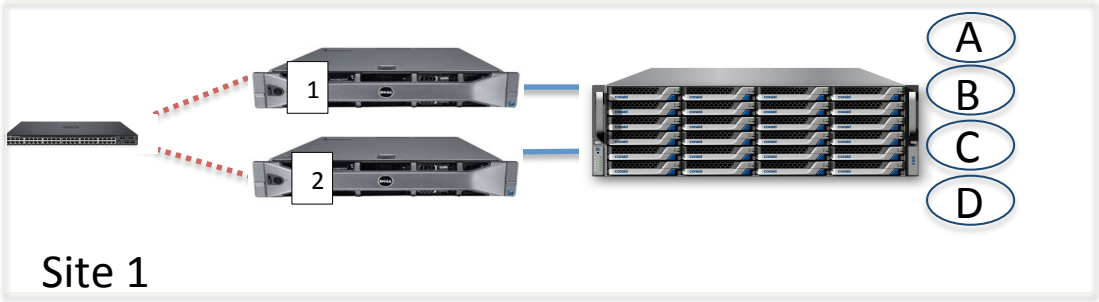
- Services are essentially OpenZFS pools (can be multiple pools per service)
- Nodes are essentially OpenZFS Appliances
- Network services are VIPs / COMSTAR views (can be multiple VIPs per service)
- Other applications/services can be added and managed

RSF-1 Stretched Cluster

HIGHAVAILABILITY

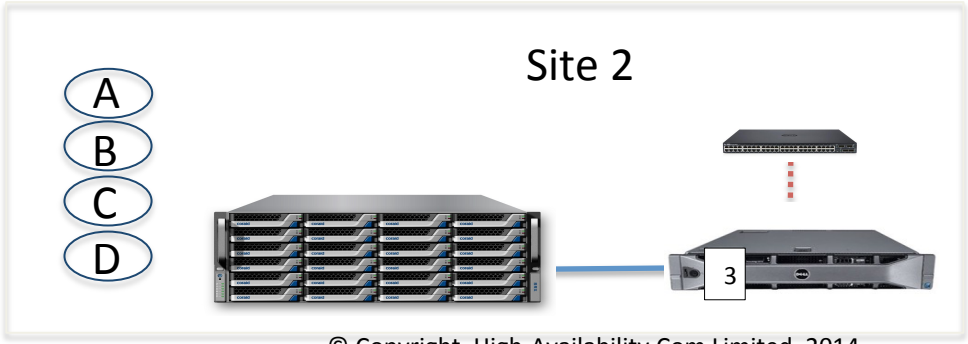
High Availability beyond the data centre

Multi-Site, Multi-Node Topology Example



Scenario:

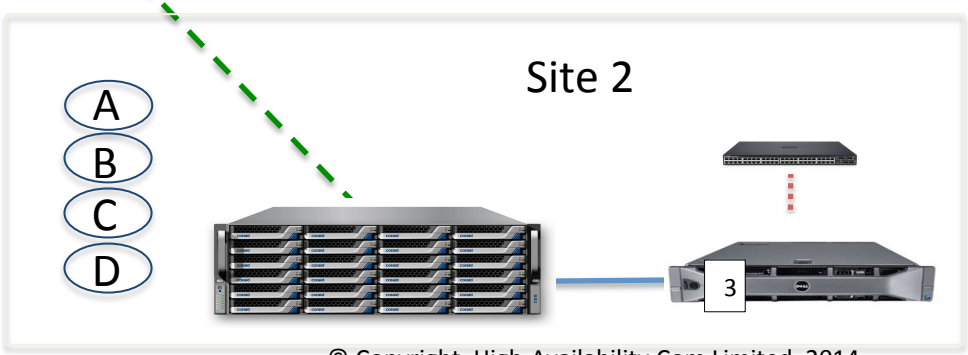
- **Servers 1 & 2 located at Site 1**
- **Server 3 located at Site 2**
- **4 ZFS Pools, A, B, C, D, each has associated VIP**



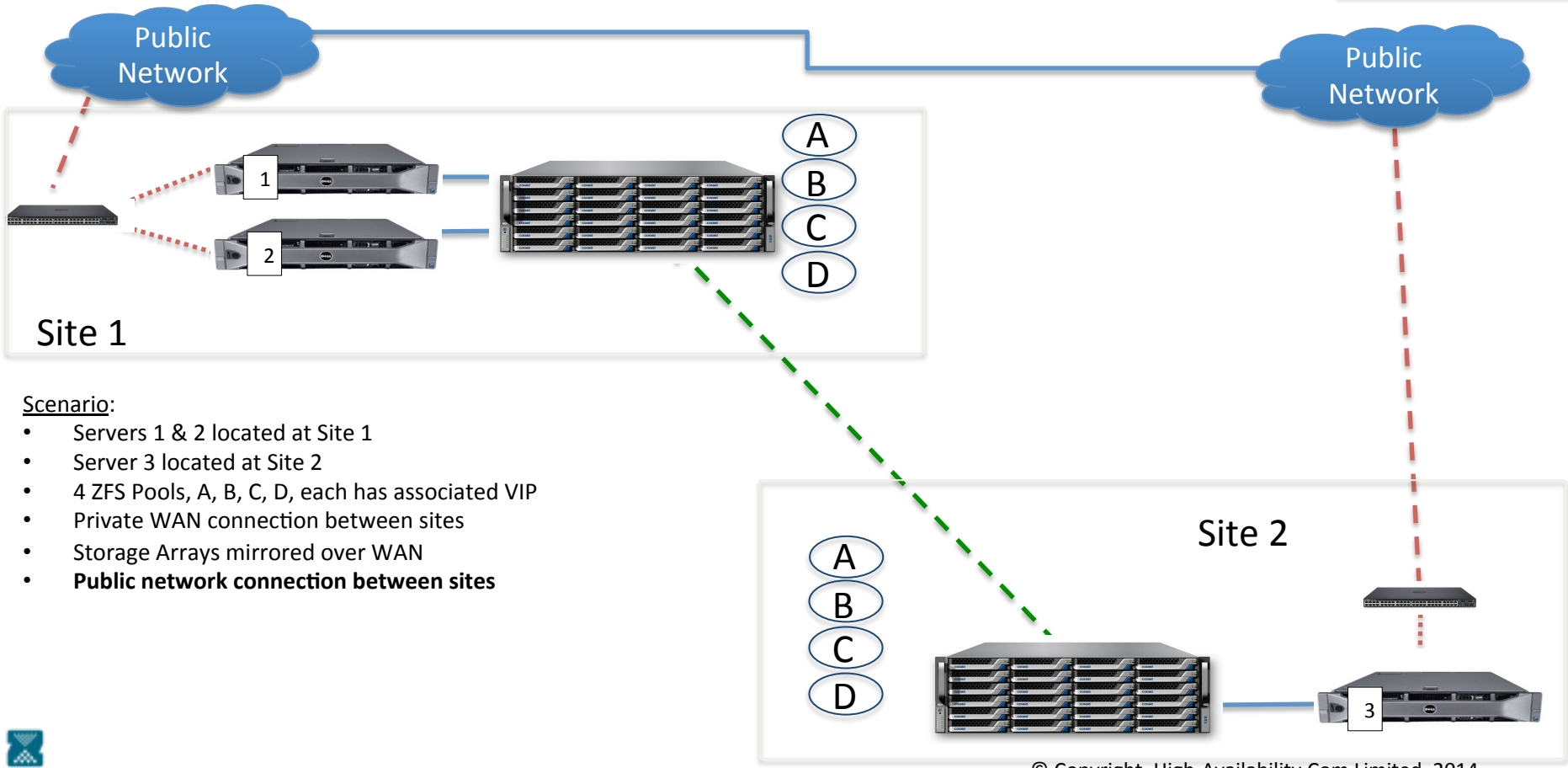
Multi-Site, Multi-Node Topology Example

Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- **Private WAN connection between sites**
- **Storage Arrays mirrored over WAN**



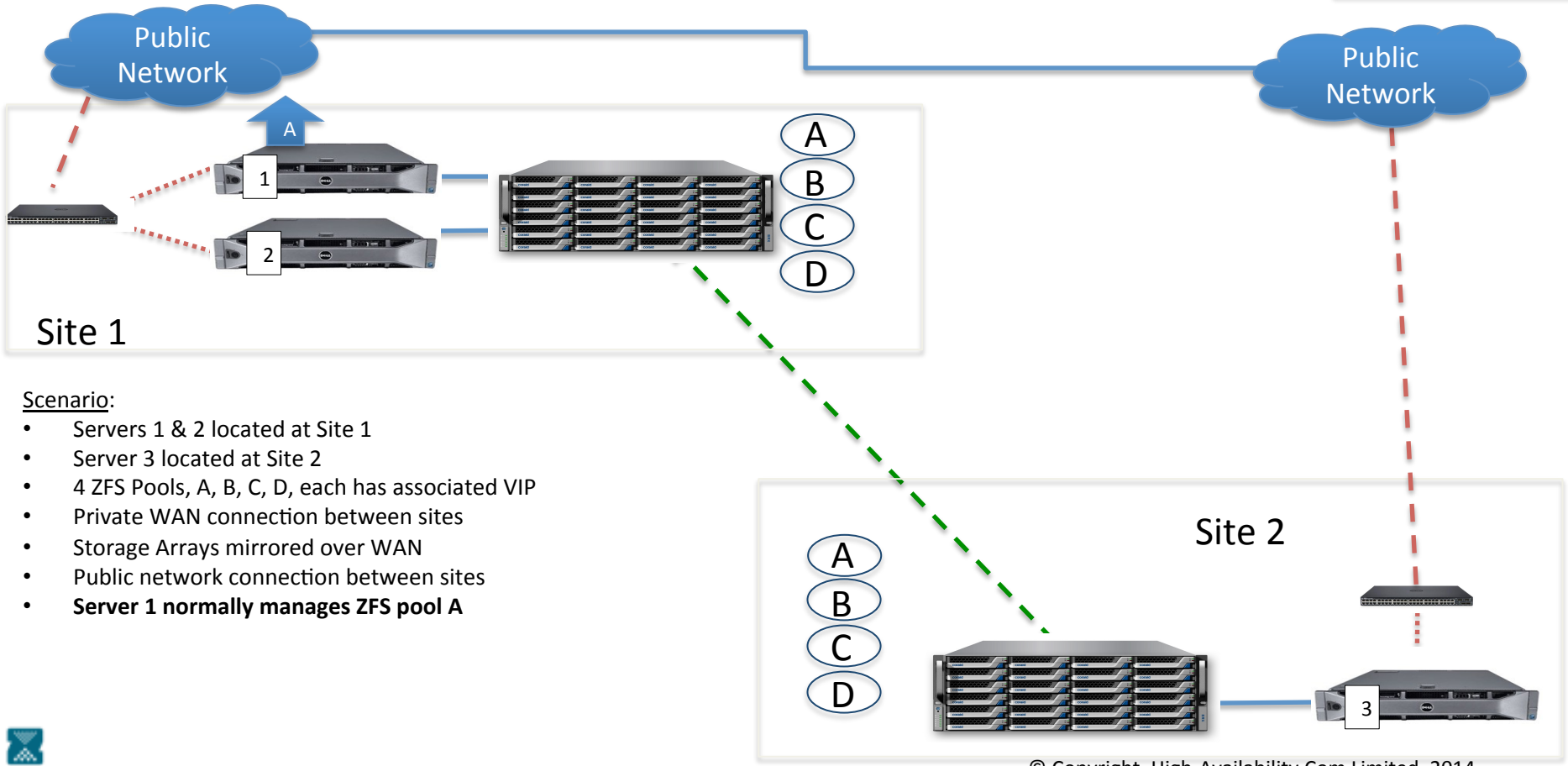
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- **Public network connection between sites**

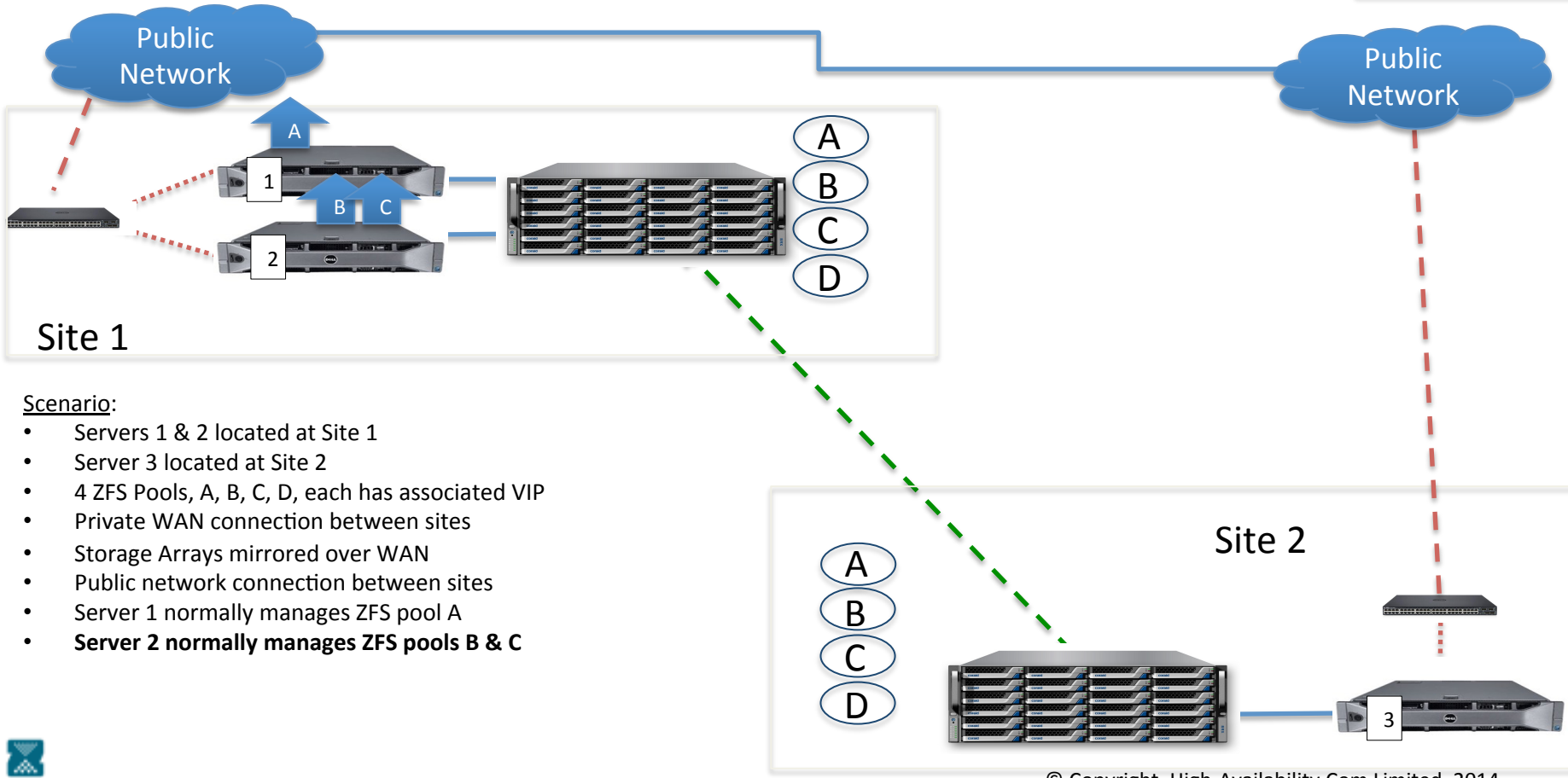
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- Public network connection between sites
- **Server 1 normally manages ZFS pool A**

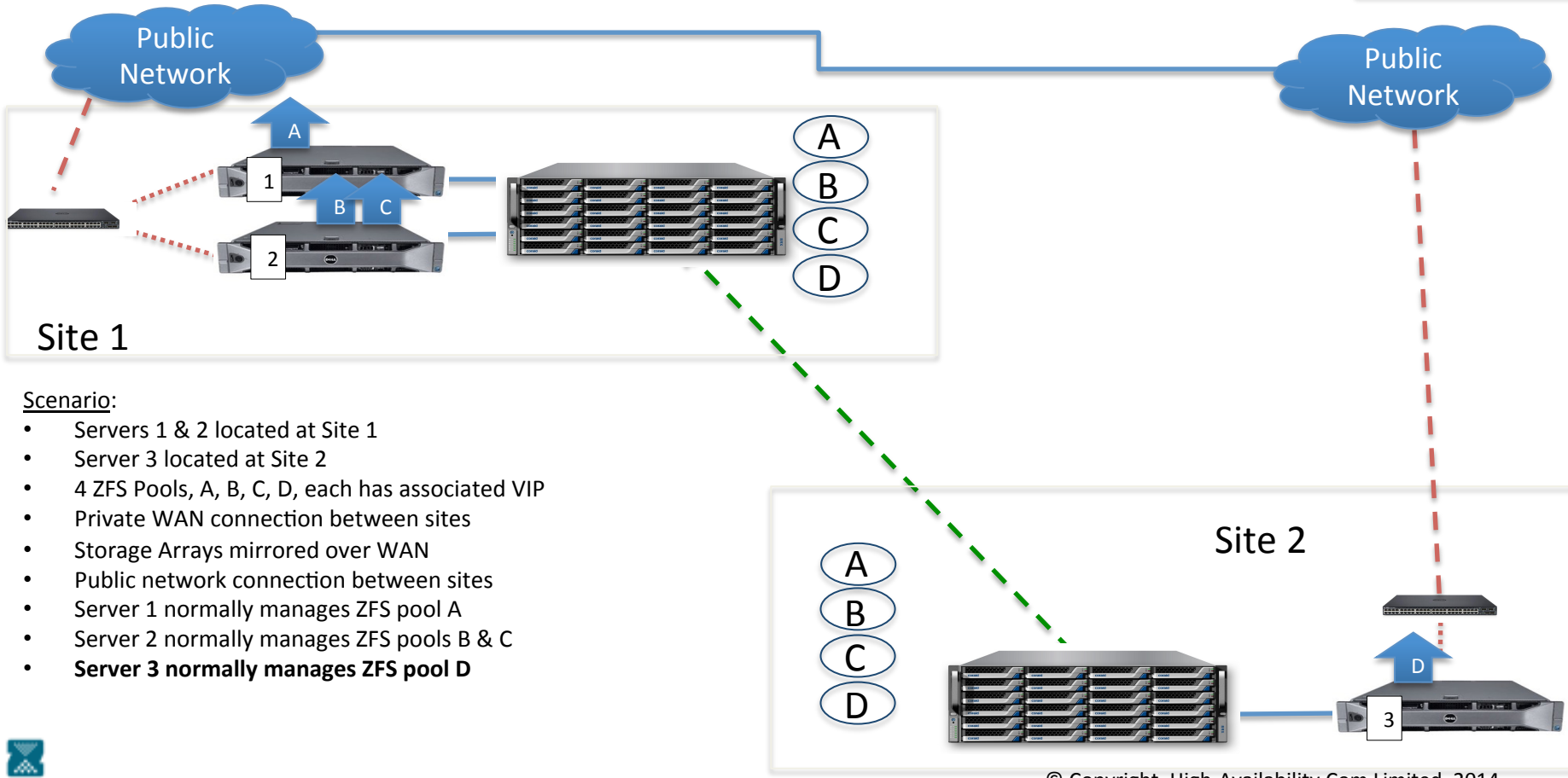
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- Public network connection between sites
- Server 1 normally manages ZFS pool A
- **Server 2 normally manages ZFS pools B & C**

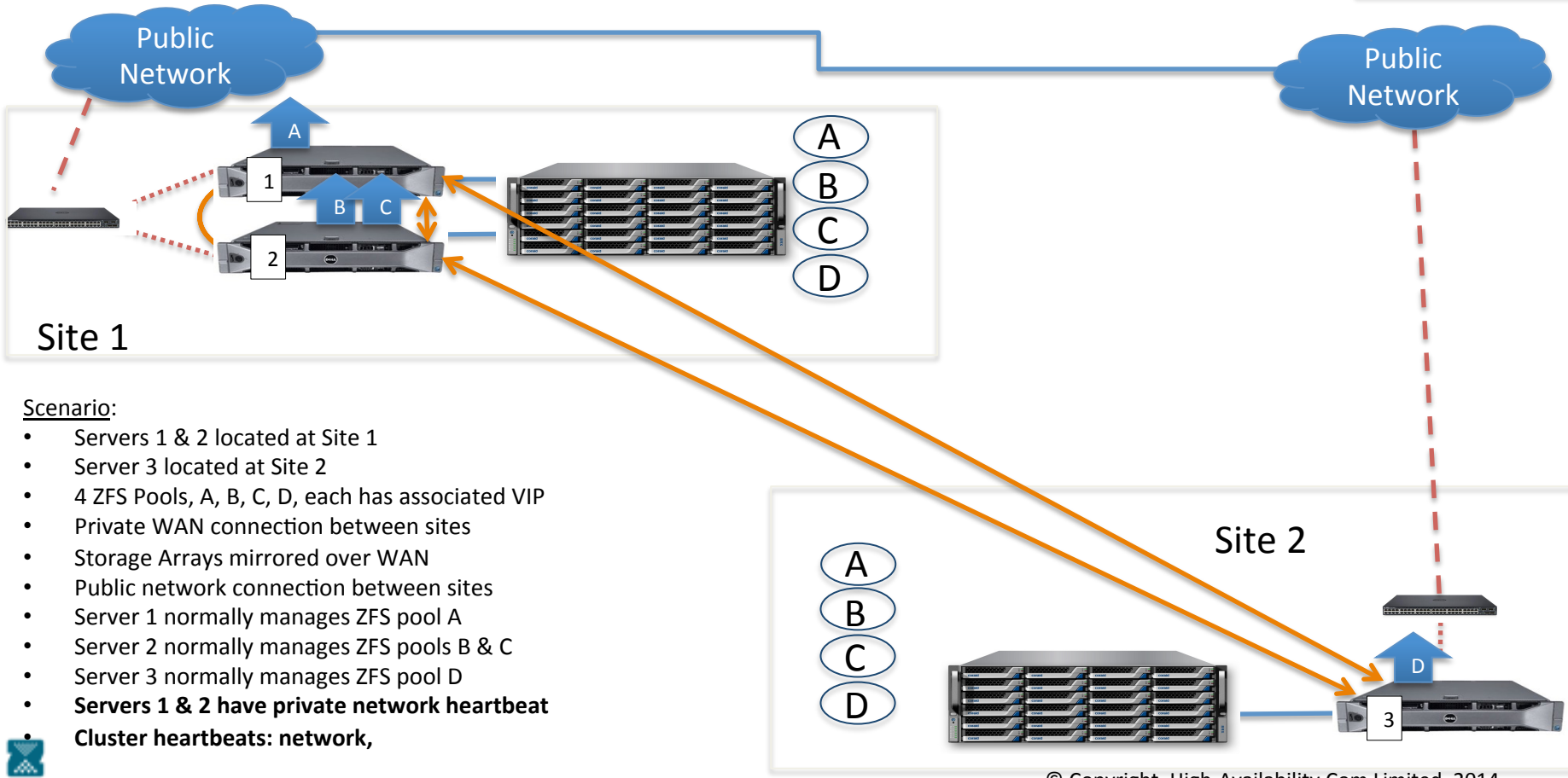
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- Public network connection between sites
- Server 1 normally manages ZFS pool A
- Server 2 normally manages ZFS pools B & C
- **Server 3 normally manages ZFS pool D**

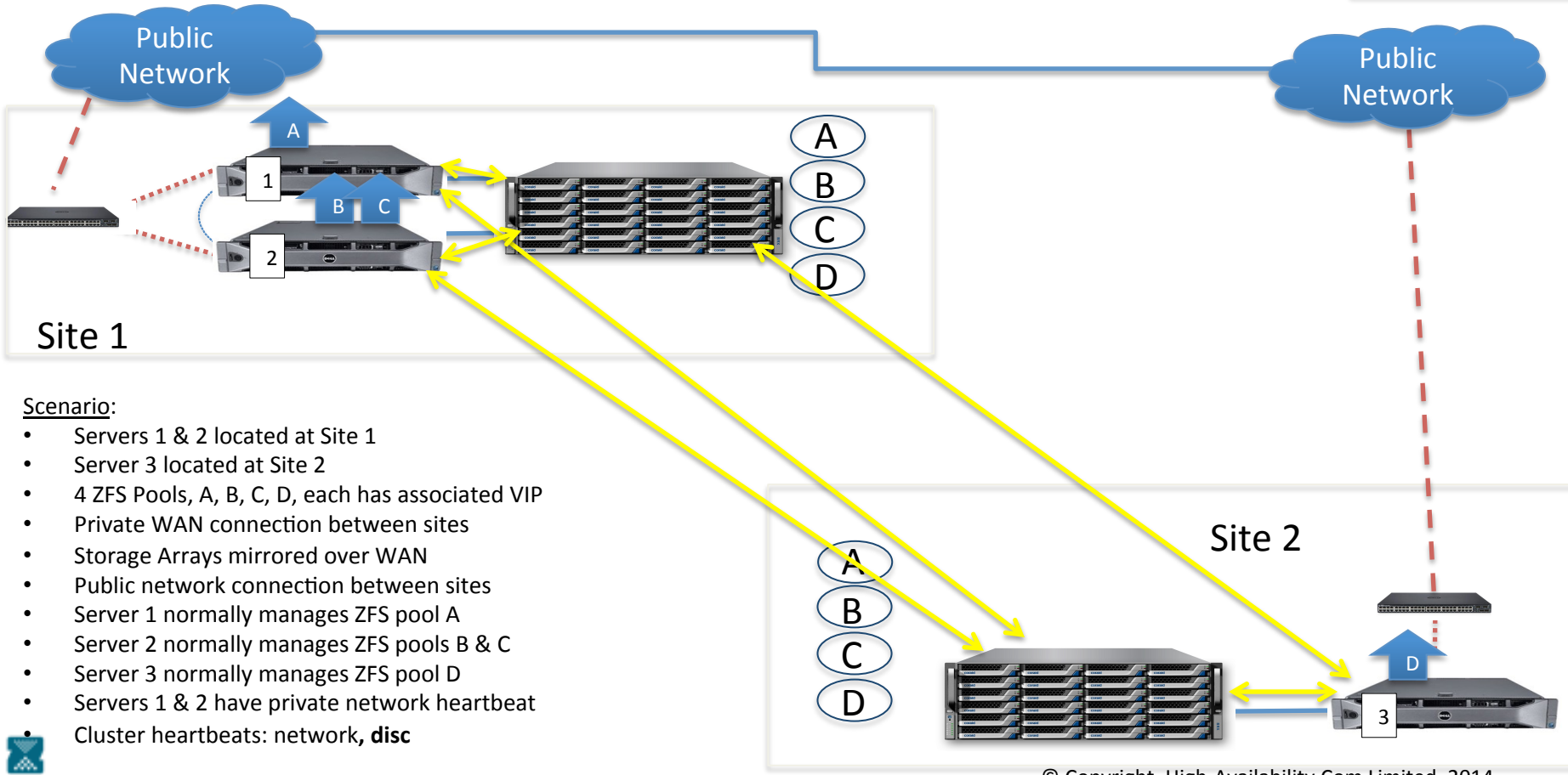
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- Public network connection between sites
- Server 1 normally manages ZFS pool A
- Server 2 normally manages ZFS pools B & C
- Server 3 normally manages ZFS pool D
- **Servers 1 & 2 have private network heartbeat**
- **Cluster heartbeats: network,**

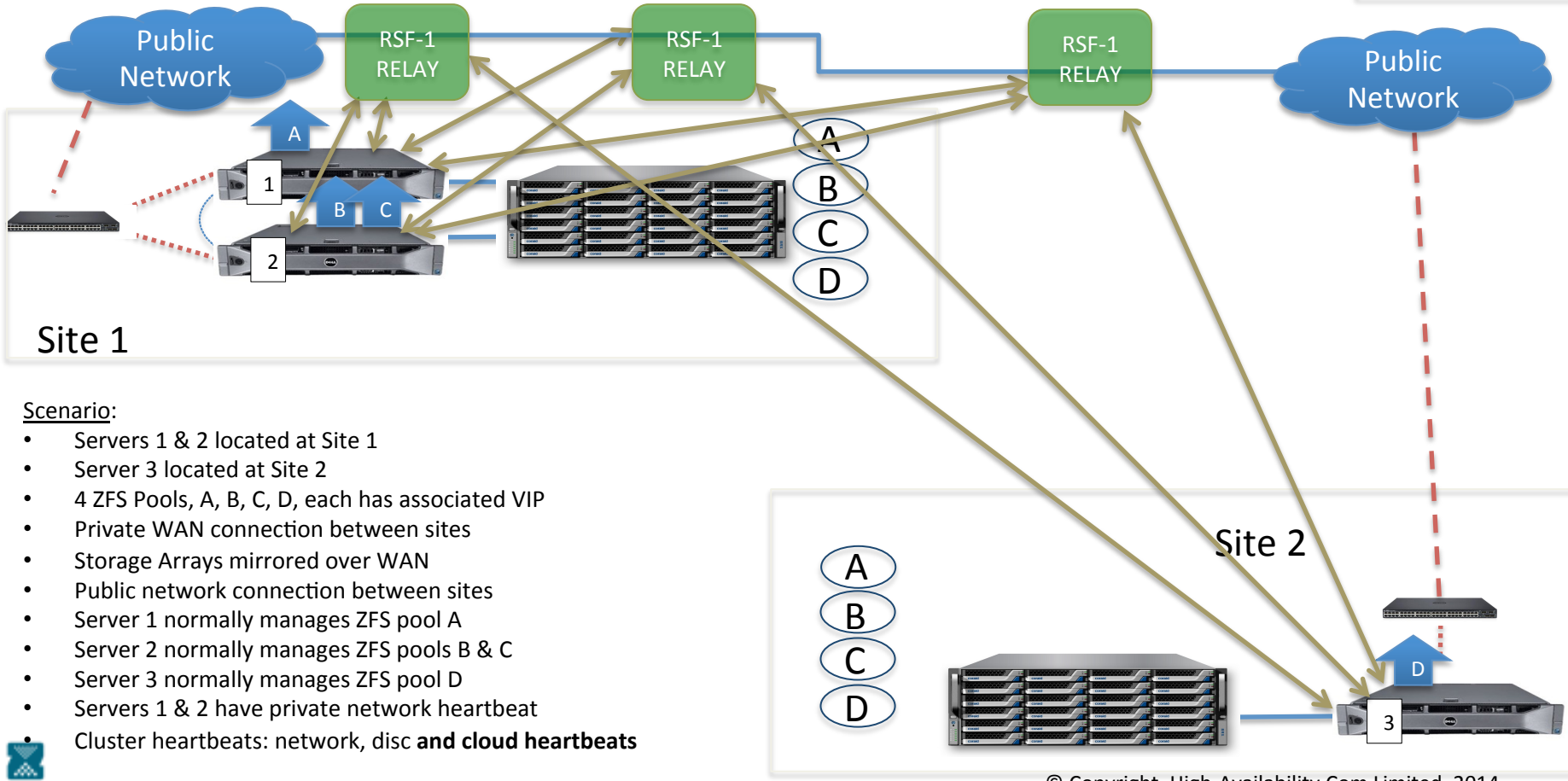
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- Public network connection between sites
- Server 1 normally manages ZFS pool A
- Server 2 normally manages ZFS pools B & C
- Server 3 normally manages ZFS pool D
- Servers 1 & 2 have private network heartbeat
- Cluster heartbeats: network**, disc**

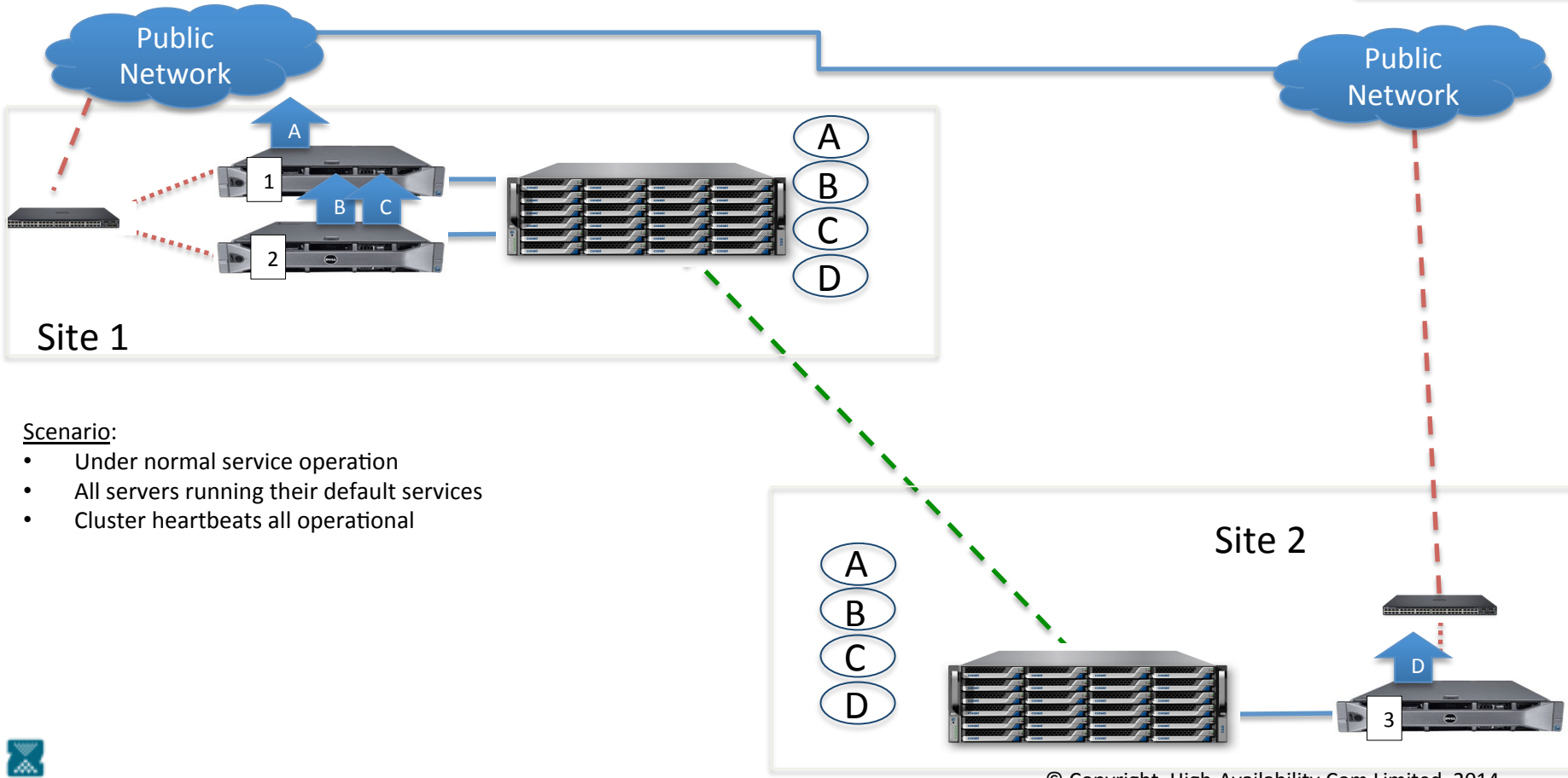
Multi-Site, Multi-Node Topology Example



Scenario:

- Servers 1 & 2 located at Site 1
- Server 3 located at Site 2
- 4 ZFS Pools, A, B, C, D, each has associated VIP
- Private WAN connection between sites
- Storage Arrays mirrored over WAN
- Public network connection between sites
- Server 1 normally manages ZFS pool A
- Server 2 normally manages ZFS pools B & C
- Server 3 normally manages ZFS pool D
- Servers 1 & 2 have private network heartbeat
- Cluster heartbeats: network, disc **and cloud heartbeats**

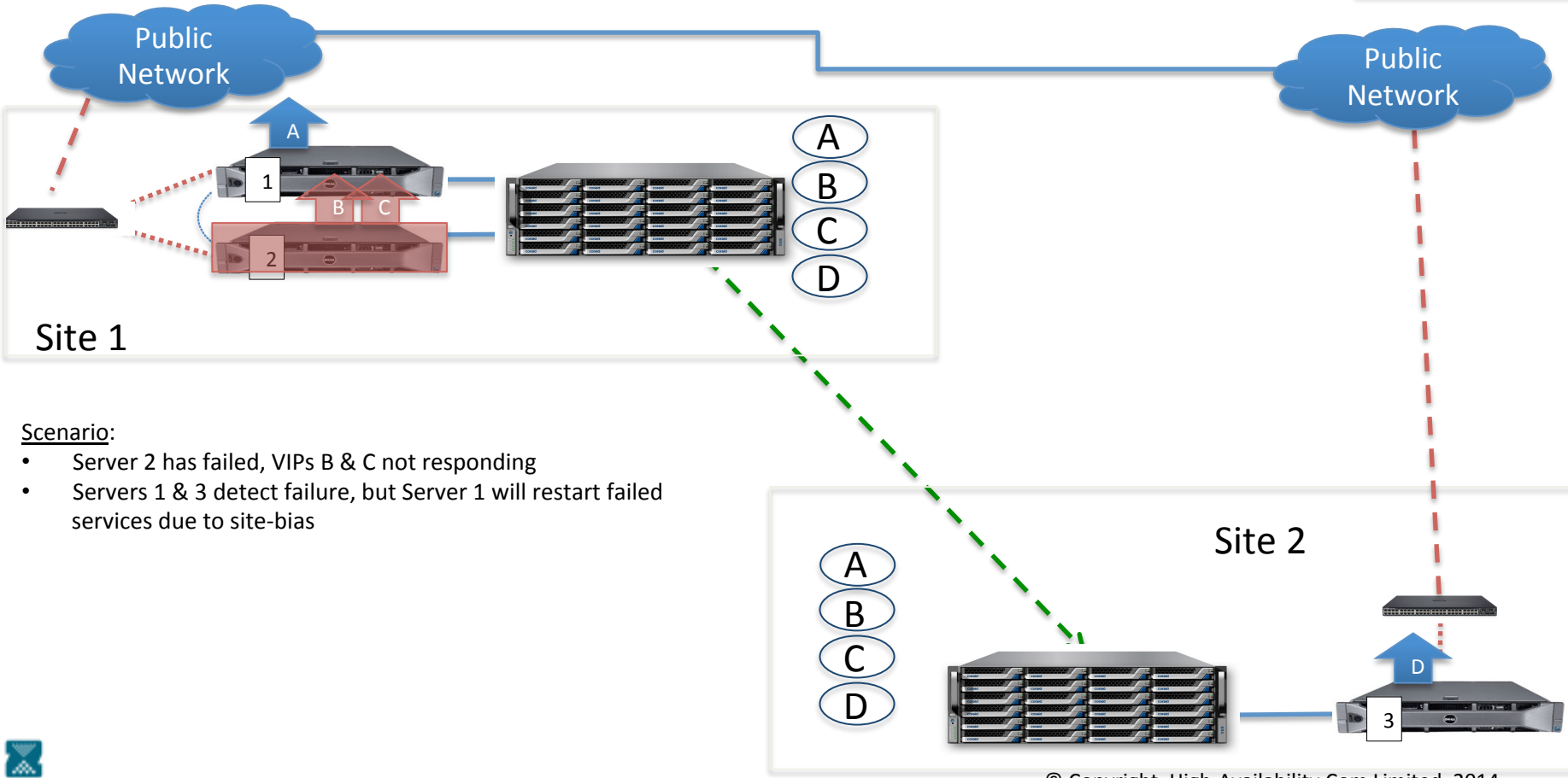
Multi-Site, Multi-Node Topology Example



Scenario:

- Under normal service operation
- All servers running their default services
- Cluster heartbeats all operational

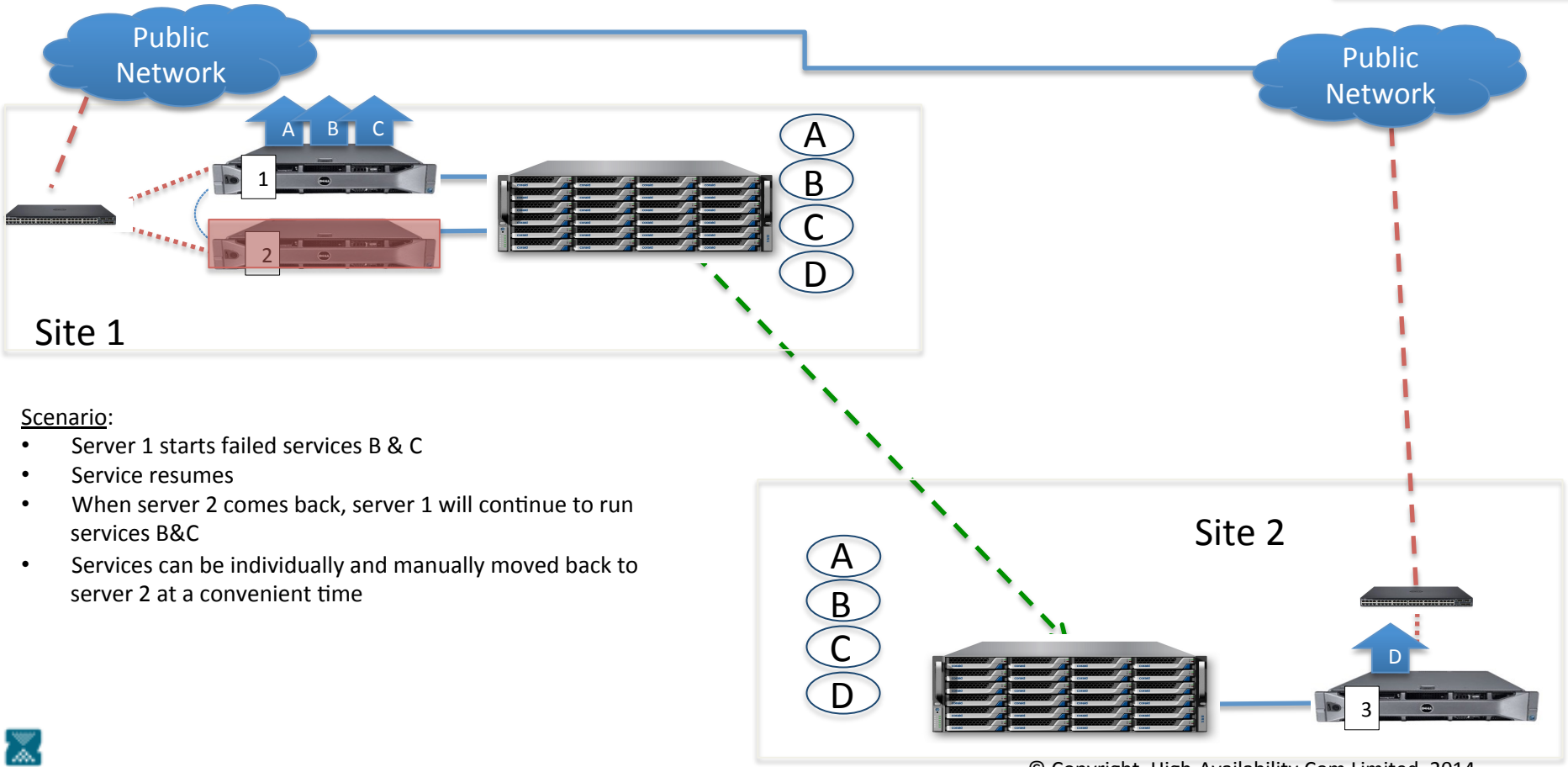
Multi-Site, Multi-Node Topology Example



Scenario:

- Server 2 has failed, VIPs B & C not responding
- Servers 1 & 3 detect failure, but Server 1 will restart failed services due to site-bias

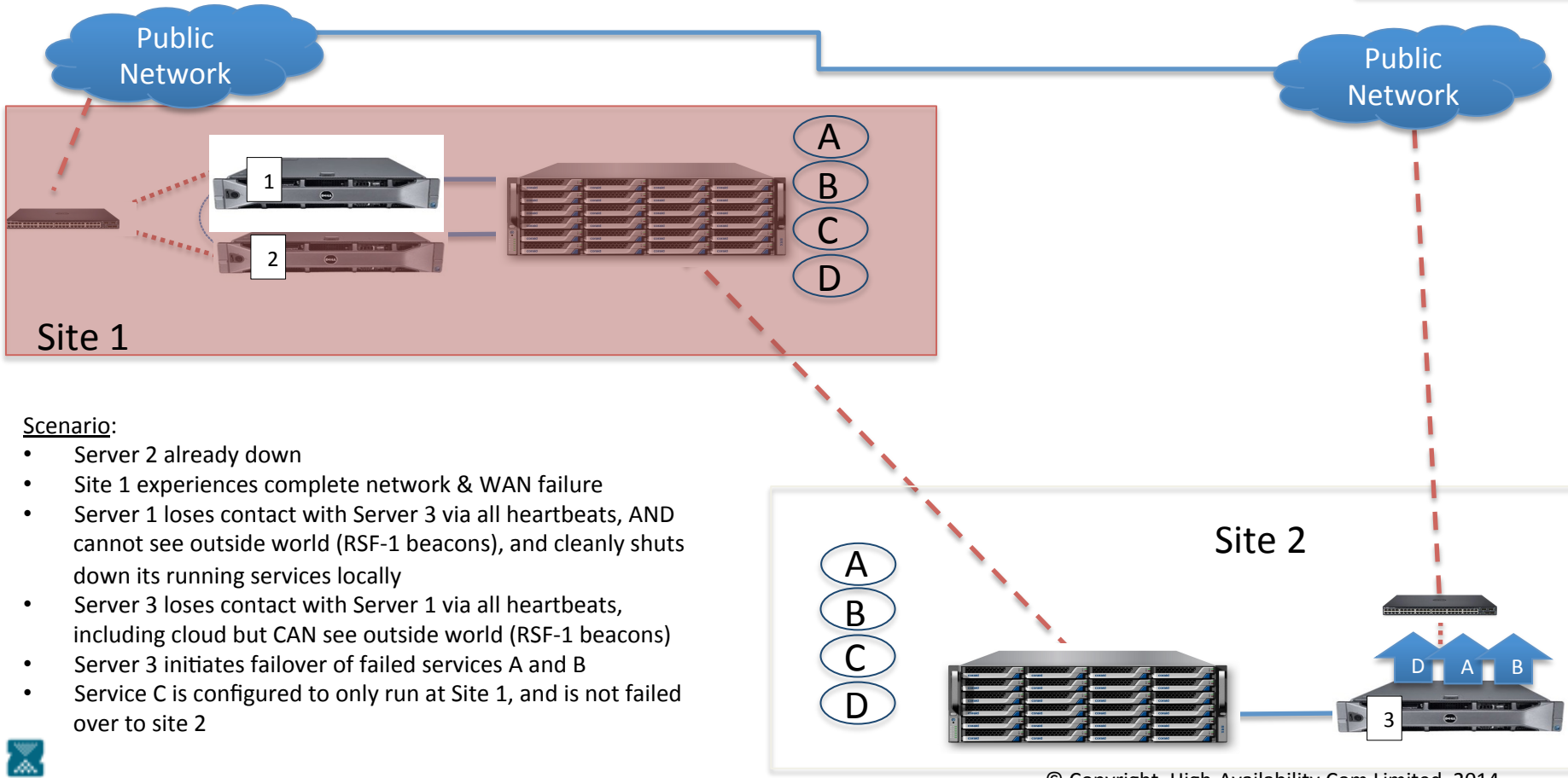
Multi-Site, Multi-Node Topology Example



Scenario:

- Server 1 starts failed services B & C
- Service resumes
- When server 2 comes back, server 1 will continue to run services B&C
- Services can be individually and manually moved back to server 2 at a convenient time

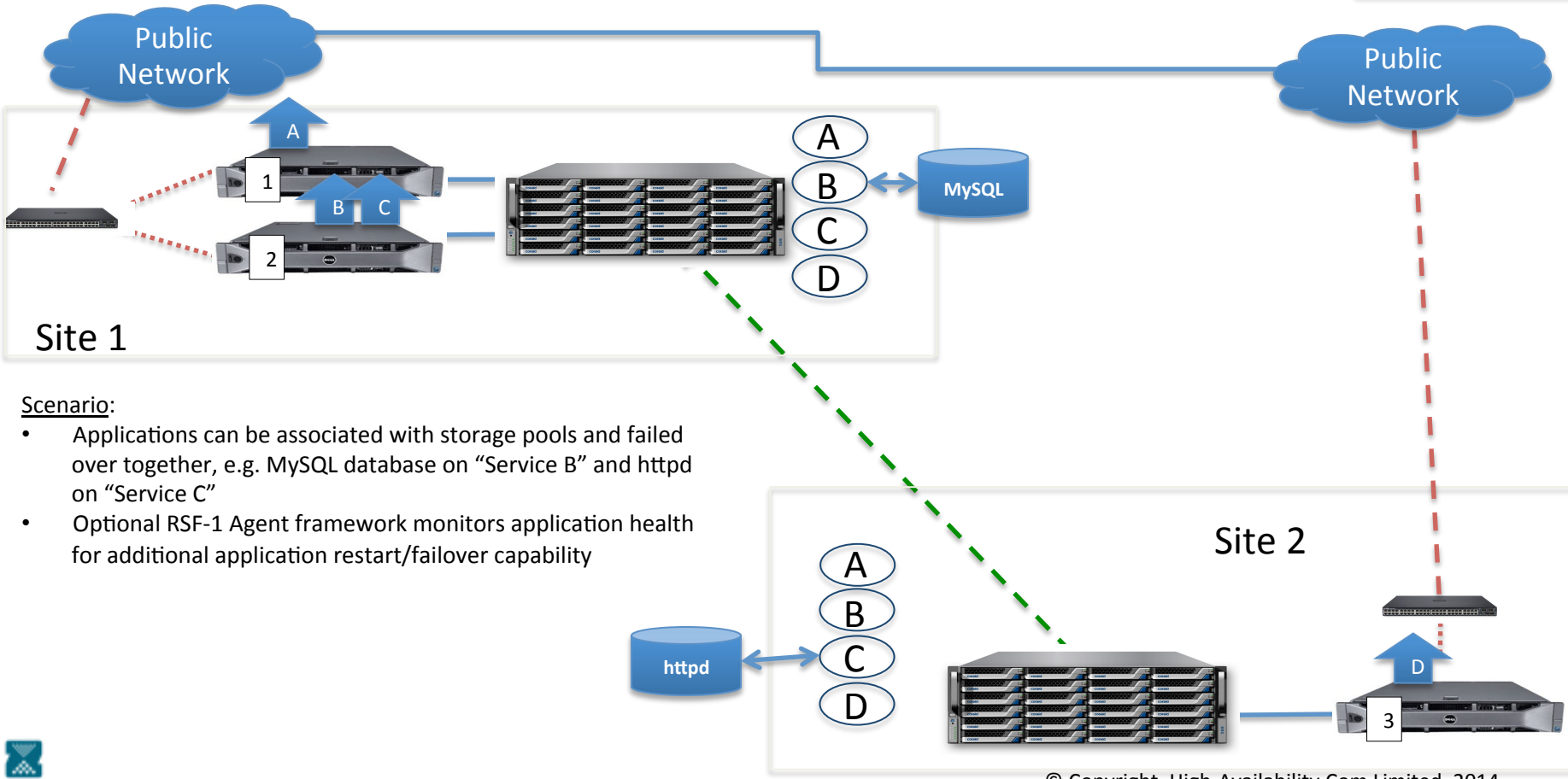
Multi-Site, Multi-Node Topology Example



Scenario:

- Server 2 already down
- Site 1 experiences complete network & WAN failure
- Server 1 loses contact with Server 3 via all heartbeats, AND cannot see outside world (RSF-1 beacons), and cleanly shuts down its running services locally
- Server 3 loses contact with Server 1 via all heartbeats, including cloud but CAN see outside world (RSF-1 beacons)
- Server 3 initiates failover of failed services A and B
- Service C is configured to only run at Site 1, and is not failed over to site 2

Application Failover



Scenario:

- Applications can be associated with storage pools and failed over together, e.g. MySQL database on “Service B” and httpd on “Service C”
- Optional RSF-1 Agent framework monitors application health for additional application restart/failover capability

RSF-1 Key Features

- Simple but Scalable
- In-Chassis, Campus Clusters, Stretch Cluster geographic topologies
- Minimal administration required
- Allows preventative maintenance without downtime
- Hot Restart - default (dynamic changes possible to the cluster)
- Active / Active / N (or Active / Passive if you like!)
- Multiple heartbeats & Multiple Types (Disk, Network, Serial, Cloud)
- Heartbeat through data disks to identify h/w issues early
- No service “bounce back” after failover
- Agent framework for software component monitoring/action
- Licence control of capability
- Platform support for Solaris, Open Solaris / Open Indiana / illumos, Linux, FreeBSD etc)
- Intelligent networking monitoring integrated with service failover and ipv6 support

RSF-1 ZFS Key Features

- Volume runs on only one appliance at a time
- Rapid Failover - initiated in seconds uses cachefile removes pool scan for vdevs
- Ease of use (2 volume cluster creation < 1 minute)
- Seamless integration with OpenZFS
- Parallel restart of pools
- Multiple Volumes & VIP's per service
- Robust data fencing mechanisms (mhdc, VIP check, heartbeat redundancy, SMITH/STONITH)
- Common Protocols & Hardware Support CIFS, NFS, iSCSI, FC
- Supports SAS, SATA (+certified interposers), FC, SCSI, AoE, iSCSI
- Client volume access via virtual IP(s) and/or COMSTAR targets
- Management of COMSTAR LU mappings/views on failover
- Full ALUA support on 2 node clusters
- Integrates support for: infiniband, ipv6, ipmi, ipmp, vlans, aggregated interfaces

Opportunities for End Users

Building your own OpenZFS Appliance?

- Add local and/or stretched HA capability
- Add HA support for any other application / service

No obligation evaluation!

- Contact us at sales@high-availability.com or visit www.high-availability.com for evaluation copies

Opportunities for Vendors

Need to add HA capability to your own offering?

- OEM / reseller /white-label options
- Seamless integration with your own user interface via comprehensive APIs
- Add HA support for any other application / service
- Flexible licensing options

No obligation evaluation!

- We'll install/configure evaluation copy on your kit
- You test
- Contact us at sales@high-availability.com or visit www.high-availability.com

Any Questions?

HIGHAVAILABILITY

www.high-availability.com

Grenville Whelan

gren@high-availability.com

OpenZFS