OpenVMS Disaster Tolerant Cluster Update

Akila B.
OpenVMS Engineering
Germany, Sep, 2009

Europe 2009 Technical Update Days

Overview

- Disaster Risks
- Disaster-Tolerant Cluster Trends
- Recent OpenVMS features – DT
- With HPVM
- Case Studies
- HP Disaster proof demo

Disaster Risks

- Some facts on the disaster risks in European countries.

- Natural disasters in the WHO European Region are increasing in number and severity. In 1990–2008, over 47 million people in the Region were affected by floods, extreme temperatures, drought, wild fires, earthquakes, accidents, mass movements (avalanches, landslides, rockfalls and subsidence) and storms. The economic damage of these events exceeded US$ 246 billion.

Source:

- http://www.euro.who.int/whd09

Health crises, excluding conflicts, in the WHO European Region, 1990–2008

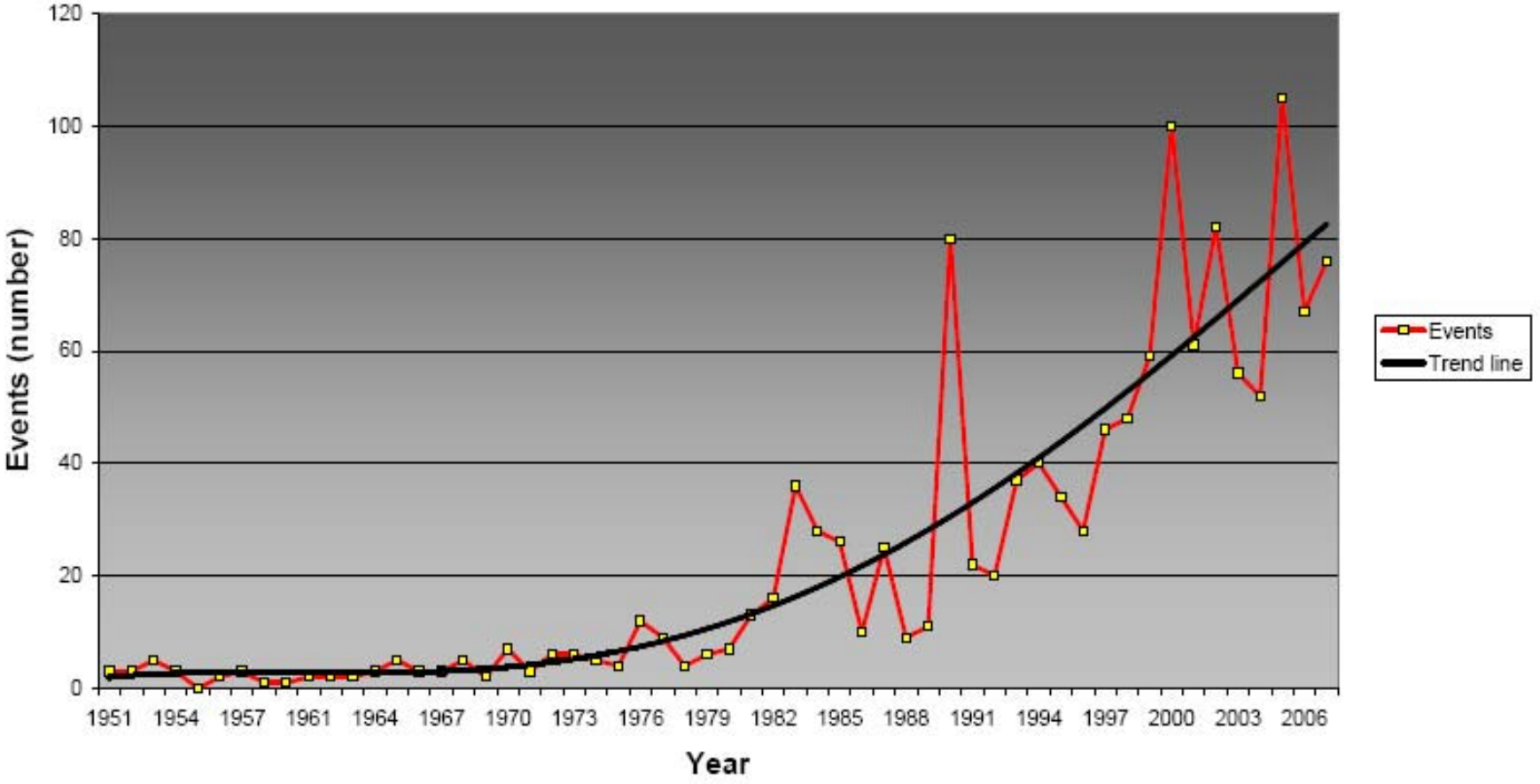
Type of event	Events (number)	Deaths (number)	People affected (number)	Economic damage (in US$ billions)
Floods	413	3 912	12 137 319	84
Extreme temperatures	141	80 993	3 442 803	16
Drought	36	2	15 875 965	15
Wild fires	72	329	1 293 432	11
Earthquakes	110	21 943	5 903 433	38
Accidents	695	18 848	154 410	12
Mass movements*	59	2 220	190 880	2
Storms	302	1 680	8 360 716	68
Total	1 828	129 927	47 358 958	246

* Mass movements include avalanches, landslides, rockfalls and subsidence.

Source: EM-DAT: Emergency Events Database [online database]. Brussels, Centre for Research on the Epidemiology of Disasters (CRED), School of Public Health, Catholic University of Louvain, 2009 (http://www.emdat.be, accessed 6 March 2009).

Fig. 1. Frequency of natural disasters in the WHO European Region, 1951–2007

Note. Natural disasters include floods, extreme temperatures, drought, wild fires, mass movements (avalanches, landslides, rockfalls and subsidence) and storms.



Source: EM-DAT: Emergency Events Database [online database]. Brussels, Centre for Research on the Epidemiology of Disasters (CRED), School of Public Health, Catholic University of Louvain, 2009 (http://www.emdat.be, accessed 6 March 2009).

Recent disasters

- Wildfires in Greece, August
- Train explosion in the Italian Riviera, June
- Earthquake in Italy, April
- H1N1 virus pandemic
- Floods in Moldova, Romania and Ukraine (summer 2008)
- Earthquake in Kyrgyzstan, October 2008
- Flooding and mudflows in Tajikistan

Sources:

- http://www.mapreport.com/subtopics/e/d.html#details
- http://www.euro.who.int/emergencies

- Disaster Risks in Europe
- **Disaster-Tolerant cluster Trends**
- Recent OpenVMS features – DT
- With HPVM
- Case Studies
- HP Disaster proof demo

Evolution OF Business Continuity

	'80s	'90s	'00s
business focus	traditional	dot.com	E-BUSINESS
requirements	restore, recover	high availability	24 x 7, scalable
recovery expectation	hardware days/hours	hardware, data minutes/seconds	hardware, data, applications minutes/seconds
decision	optional		mandatory

Today Business Continuity plan is key regulatory and legal requirement in many geographies/industry verticals

Disaster-tolerant cluster trends

- Distance Trends:
  - Longer inter-site distances for better protection (or because the customer already owns datacenter sites in certain locations)
  - Business pressures for shorter distances for better performance

Disaster-tolerant cluster trends

- Network Trends:
  - Inter-site links getting cheaper and higher in bandwidth
  - Harder to get dark fiber; easier to get lambdas (DWDM channels)
  - Ethernet of various speeds is available for cluster interconnects
  - IP network focus; increasing pressure not to bridge LANs between sites
  - Inter-site links:
    - DS-3 [E3 in Europe],
    - OC-3, - Preferred for WANs
    - OC-12,
    - OC-48,
    - Dark fiber, - Sites sharing same campus
    - Lambdas (individual channels over DWDM) – preferred for MANs

18-Sep-09

Disaster-tolerant cluster trends

- Storage Trends
  - Bigger, faster, cheaper disks; more data needing replication between sites
  - Faster storage area networks
  - Storage:
    - Local SCSI or SAS/SATA disks, or
    - Fibre Channel (100km), SAN-based storage
  - Inter-site SAN links:
    - Direct fiber-optic links for short distances
    - SAN Extension using Fibre Channel over IP (FCIP) for longer distances

Hardware Trend

- Servers – Existing clusters
- Blades – New installations
- Virtual Machines – Future option for DT

General questions - DT

- Multi-site cluster or just data replication (using Remote Vaulting)?
- 2-site or 3-site cluster?
- Quorum scheme (balanced votes, disk, node)
- Distance between sites?
- Performance Vs Distance (1 ms per 50 miles)

Questions…

- Continuous Access or Host Based Volume shadowing?
- FC connection, still MSCP required?
- Shadow System disk across sites?

- Disaster Risks in Europe
- Trends
- **Recent OpenVMS features – DT**
- With HPVM
- Case Studies
- HP Disaster proof demo

- OpenVMS Cluster Features in Response to customer needs on DT:

➢Host-Based Volume Shadowing

✓AMCVP

✓6 Member shadowset

➢Cluster Communication

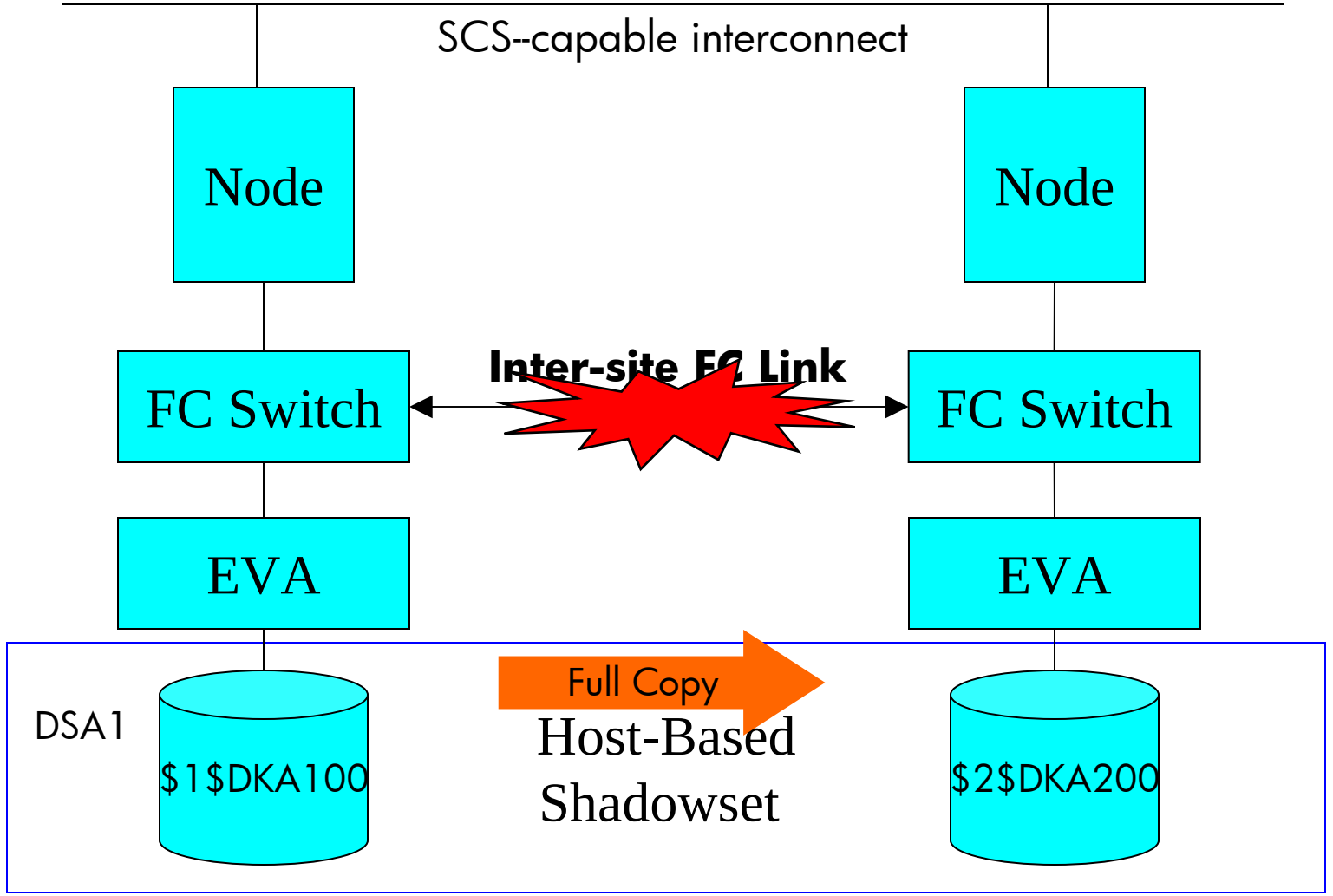
✓IPCI (IP as Cluster Interconnect)

➢HPVM

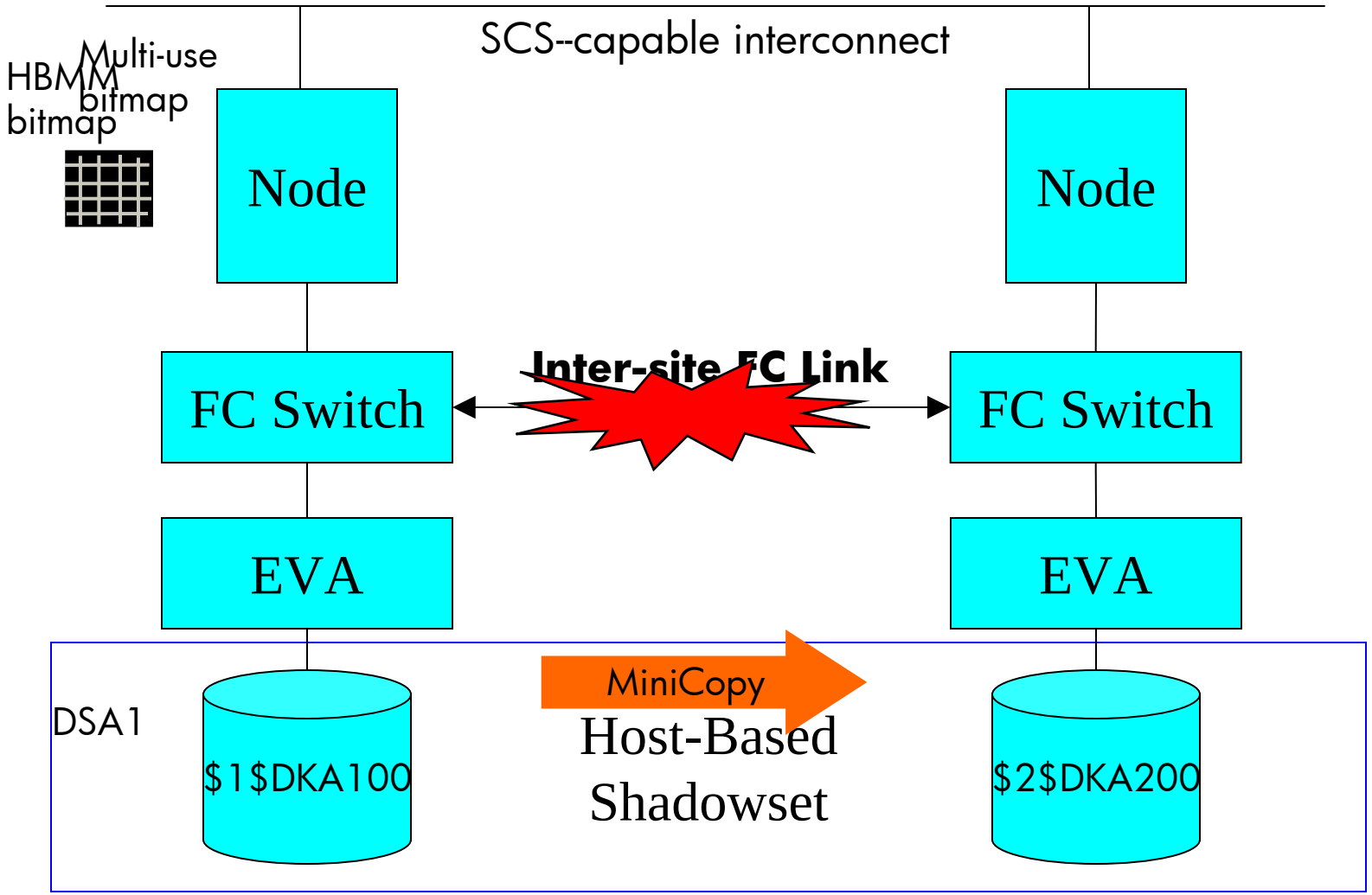
Scenario 1

- Temporary Site outage
- A shadow member is removed due to unexpected failure, like inter-site link failure.
- Full copy is required when the member is added back to the shadowset.
- OpenVMS V8.3 allows Mini-Merge bitmaps to be converted to Mini-Copy bitmaps for **quick recovery** from unscheduled site outage

Full copy when member expelled



Automatic Minicopy on Volume Processing (AMCVP)

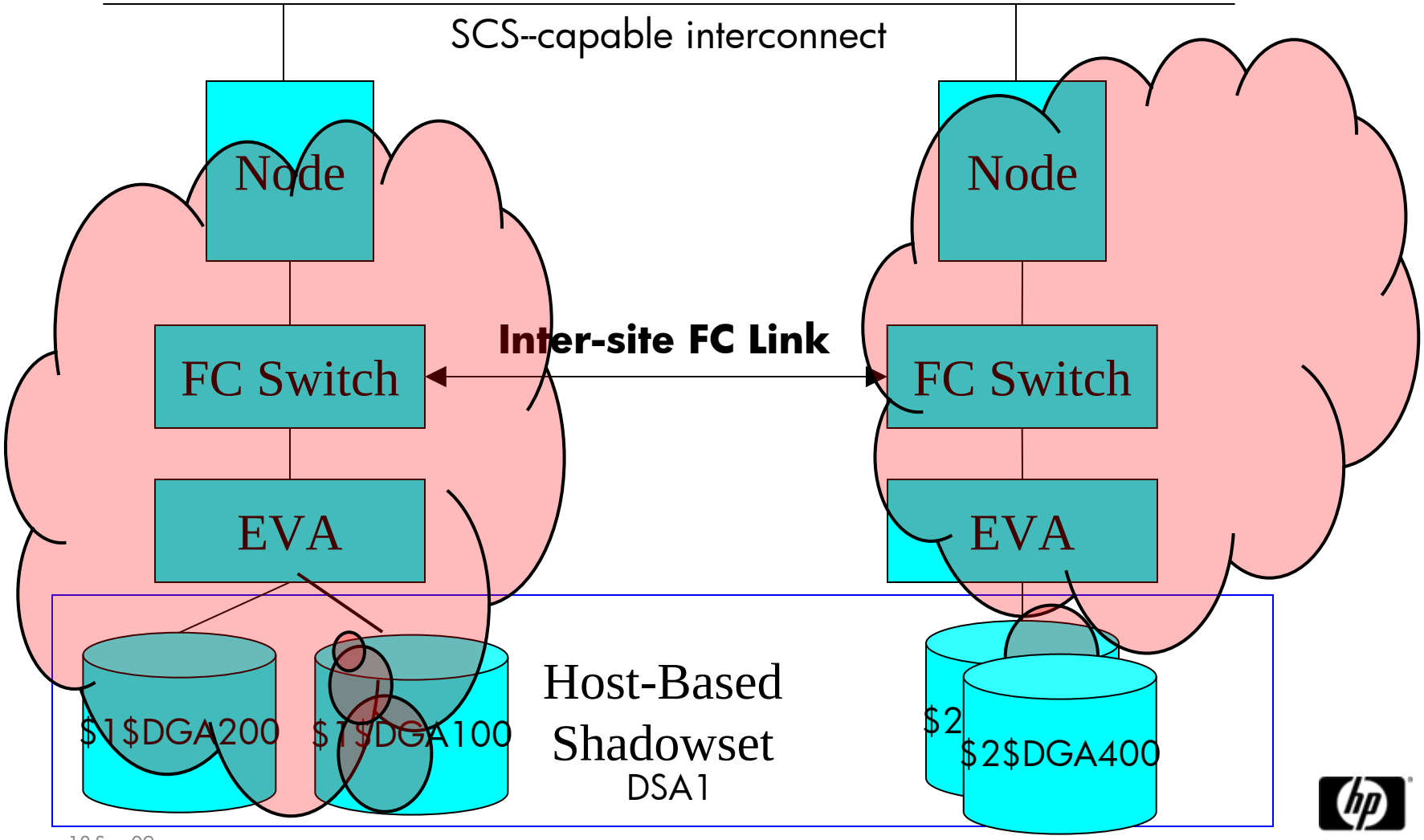


Scenario 2

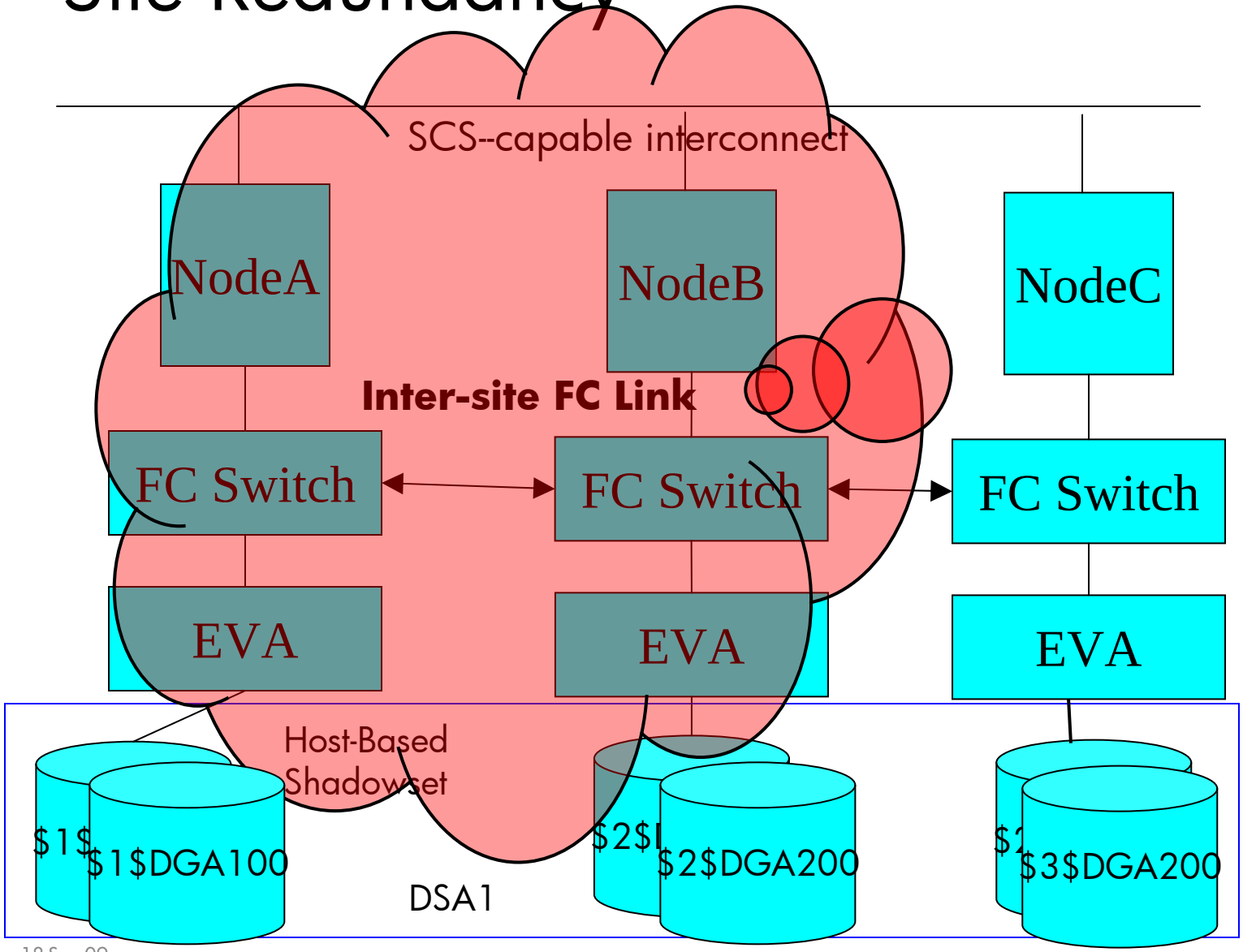
- Desire to still have redundancy of storage after a site failure in a disaster-tolerant cluster

- OpenVMS next release to support up to 6-member shadowsets compared with the current limit of 3-member.

2 Site Redundancy



3 –Site Redundancy



Scenario 3

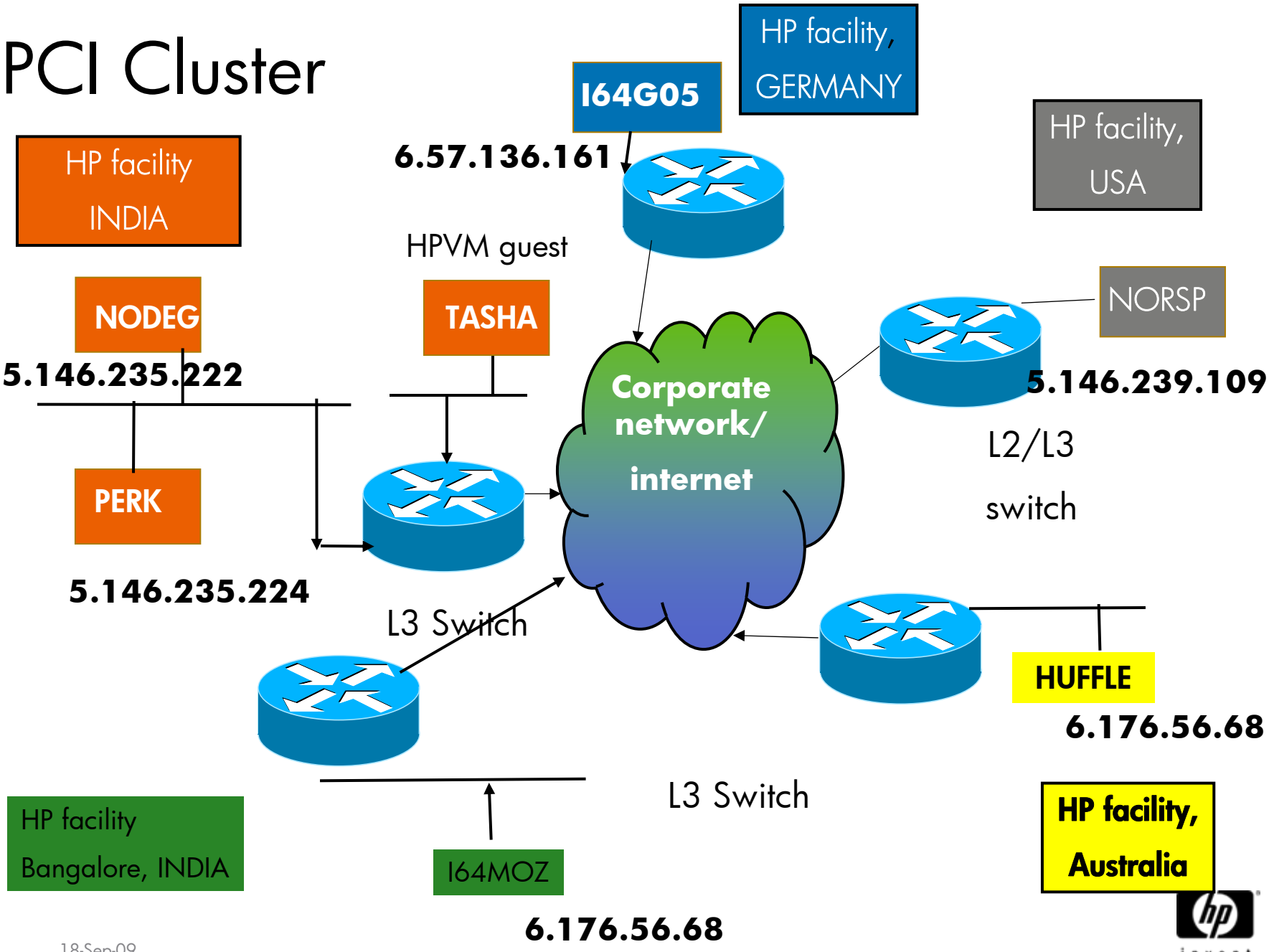
- My network folks refuse to allow any bridging between sites. How can I set up a multi-site OpenVMS disaster-tolerant cluster?
- IP routing as network preferred, not bridging

- OpenVMS next release to support IP as a Cluster Interconnect

- IP as a Cluster Interconnect will help raise 3-site numbers significantly, as it will make having a quorum node at a 3rd site much easier

IPCI Cluster

IPCI Cluster

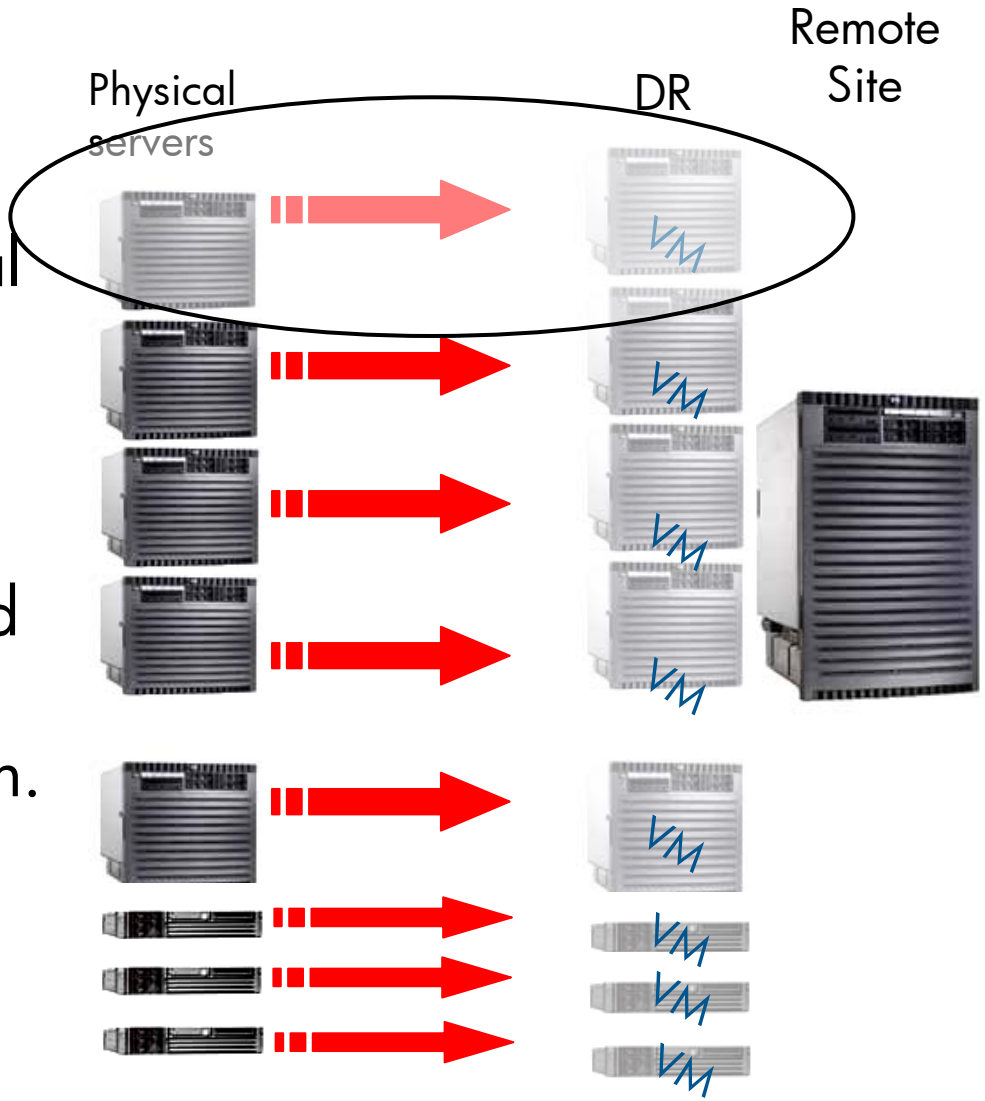


OpenVMS as guest on HPVM

- Single hardware can support multiple OpenVMS Virtual machines as guest.
- OpenVMS Cluster support should be available in next release.
- Cluster can be formed between Virtual and physical nodes.
- Beneficial when need to setup failover site for large number of nodes spread across.

HPVM - DT

- Benefits
  - Redundant virtual servers
  - Fewer physical servers at DT site
  - VMs ready-to-boot/active standby; otherwise used for development, test, evaluation.



- Disaster Risks in Europe
- Trends
- Recent OpenVMS features – DT
- With HPVM
- **Case Studies (6)**
- HP Disaster proof demo

Case Study 1: Global Cluster

- Internal HP test cluster
  - IP network as Cluster Interconnect
- Sites in India, USA, Germany, Australia
- Inter-site distance (India-to-USA) about 8,000 miles
  - Estimated circuit path length about **22,000** miles
- Round-trip latency of about **350** milliseconds

Case Study 2: 3,000-mile Cluster

- **3,000** mile site separation distance
- Disaster-tolerant OpenVMS Cluster
- Originally VAX-based, thus running for many years now, so presumably acceptable performance

Case Study 3: HBVS over CA

- Short-distance 3-site OpenVMS disaster-tolerant cluster configuration.
- 3rd site - quorum node. Moved from quorum disk on an EVA at one of the two sites, to 3rd site quorum node.
- Inter-site link(s): 1 Gbit Fibre Channel, multiple 1-gigabit Ethernet LAN connections
- Storage: Storageworks EVA at each of two sites
- Customer compared **HBVS and Continuous Access** (CA) and chose HBVS because he couldn't get CA to fail over in faster than 2 minutes

Case Study 4: Proposed 600-mile Cluster

- Existing OpenVMS DT cluster with 1-mile distance
- One of two existing datacenters is to be closed
- Proposed moving one-half of the cluster to an existing datacenter 600 miles away
  - Estimated circuit path length about **800** miles
  - Round-trip latency **13** milliseconds

Case Study 4:

- Month-end processing time is one of the most performance-critical tasks
- Tested in OpenVMS Customer Lab using D4
- Performance impact high.
- May do shorter-distance DT cluster to new site, then use CA (Asynchronous) to distant site for DR purposes

Case Study 5: Proposed DT Clusters using HPVM

- Educational customer, state-wide network
- OpenVMS systems at 29 remote sites
- Proposed using HPVM on Blade hardware and storage at central site to provide 2nd site and form disaster-tolerant clusters for 29 other sites simultaneously
- Most of the time only Volume Shadowing would be done to central site
- Upon failure of any of the 29 sites, the OpenVMS node/instance at the central site would take over processing for that site

Special case: 20-mile DT Cluster

- Existing OpenVMS Cluster
- Needed protection against disasters
- Implemented DT cluster to site 20 miles away
  - Estimated circuit path length about **50** miles
  - **0.8** millisecond round-trip latency

Special case : 20-mile DT Cluster

- Performance of night-time batch jobs had been problematic in the past
  - CPU saturation, disk fragmentation, directory files of 3K-5K blocks in size, and need for database optimization were potential factors
  - After implementing DT cluster, overnight batch jobs now took hours too long to complete
    - **Slower write latencies** identified as the major factor
    - Former factors still uncorrected

Special case : 20-mile DT Cluster Write Latencies

- MSCP-serving is used for access to disks at remote site. Theory predicts **writes take 2 round trips**.
- Write latency to local disk measured at 0.4 milliseconds
  - Write latency to remote disks calculated as:
    - 0.4 + (twice 0.8 millisecond round-trip time) = 2.0 milliseconds
  - Factor of **5X** slower write latency

Special case : 20-mile DT Cluster Write Latencies

- FCIP-based SAN Extension with Cisco *Write Acceleration* or Brocade *FastWrite* would allow writes in one round-trip instead of 2
  - Write latency to remote disks calculated as:
    - 0.4 + (once 0.8 millisecond round-trip time) = 1.2 milliseconds
  - Factor of **3X** slower write latency instead of 5X

Special case : 20-mile DT Cluster Read Latencies

- Disk selected for read:
  - local queue length of device + Read_Cost
  - Lowest total_read_cost member selected.

Default OpenVMS Read_Cost values:

  - Local Fibre Channel disks = 2
  - MSCP-served disks = 501
    - Difference of 499

- If Queue length at local site = 499 or above, then MSCP path is used for read

Special case : 20-mile DT Cluster Read Latencies

- Read latency to remote disks calculated as:
  - 0.4 + (one 0.8 millisecond round-trip time for MSCP-served reads) = 1.2 milliseconds
  - 1.2 milliseconds divided by 0.4 milliseconds is 3
  - At a local queue length of 3 you get a response time equal to the remote response time, so certainly at a local queue depth of 4 or more it might be beneficial to start sending some of the reads to the remote site
  - Difference in Read_Cost values of around 4 might work well

Special case : 20-mile DT Cluster

- Presently remove remote shadowset members each evening to get acceptable performance overnight, and put them back in with Mini-Copy operations each morning.
  - Recovery after a failure of the main site would include re-running night-time work from the copy of data at the remote site
  - Business requirements in terms of RPO, RTO happen to be lenient enough to permit this strategy

How to move a system with no downtime

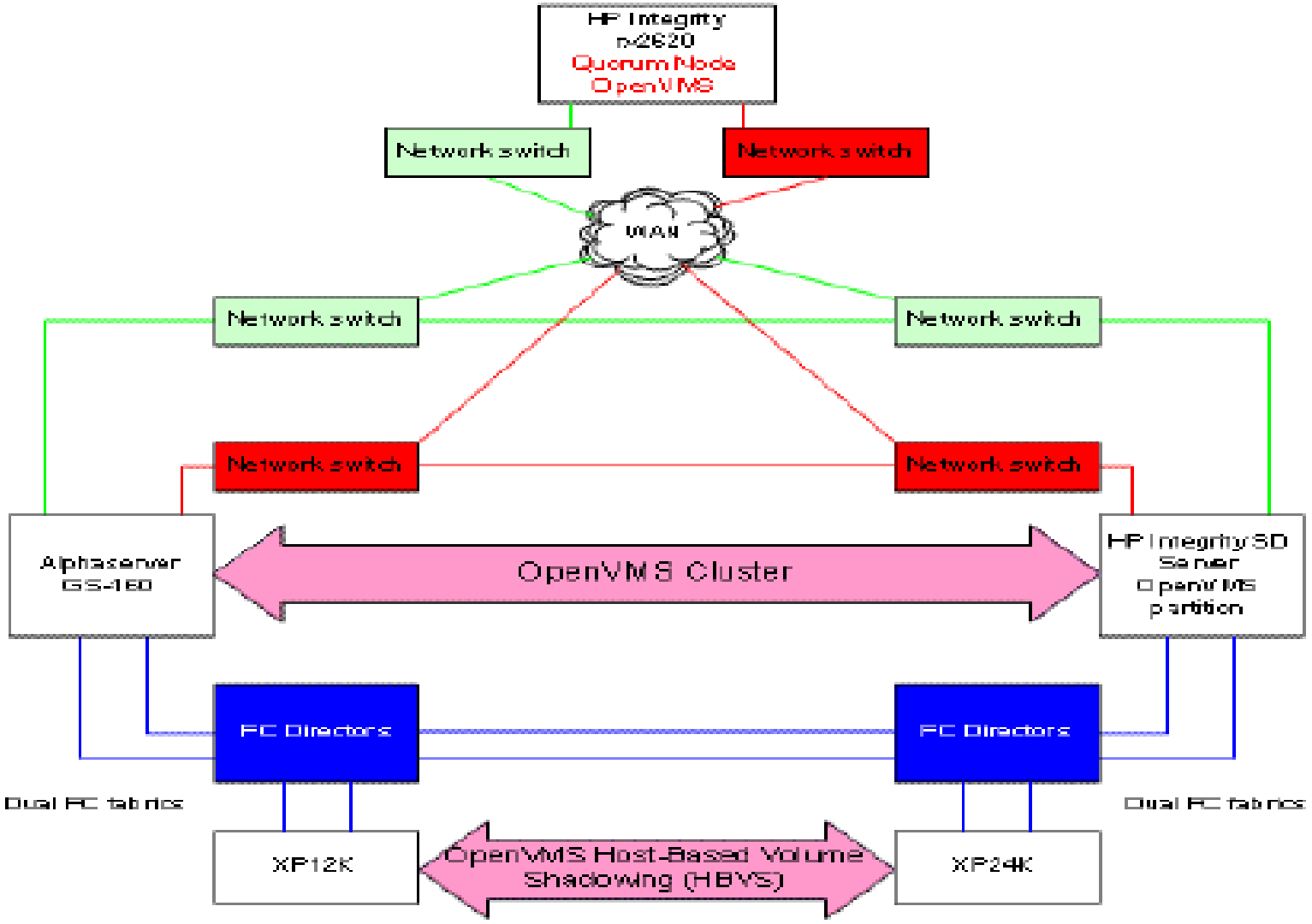
- Disaster Risks in Europe
- Trends
- Recent OpenVMS features – DT
- With HPVM
- Case Studies
- **HP Disaster proof demo**

HP Disaster recovery test

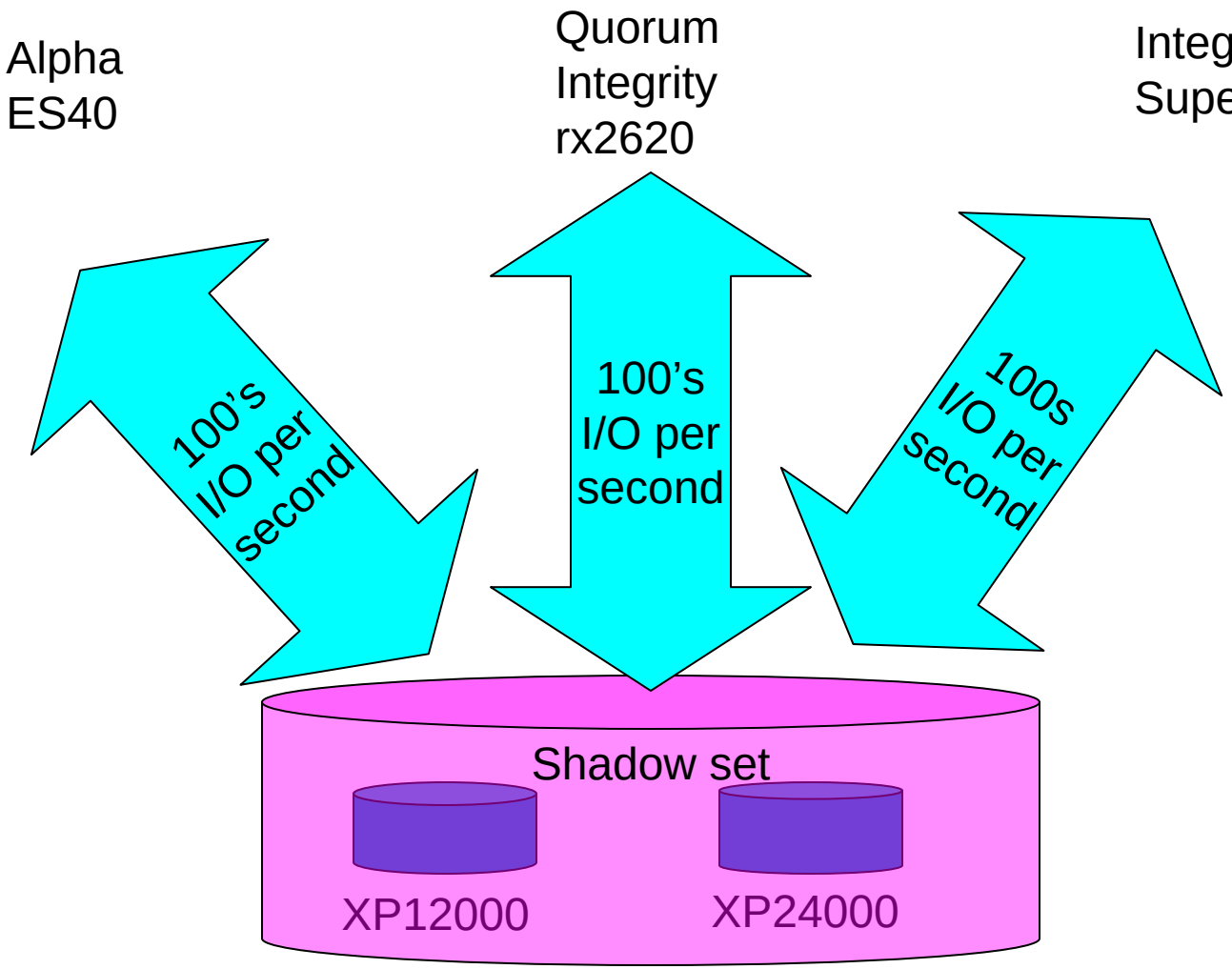
Original “green” datacenter

Failover datacenter

Disaster Proof Demo OpenVMS Cluster

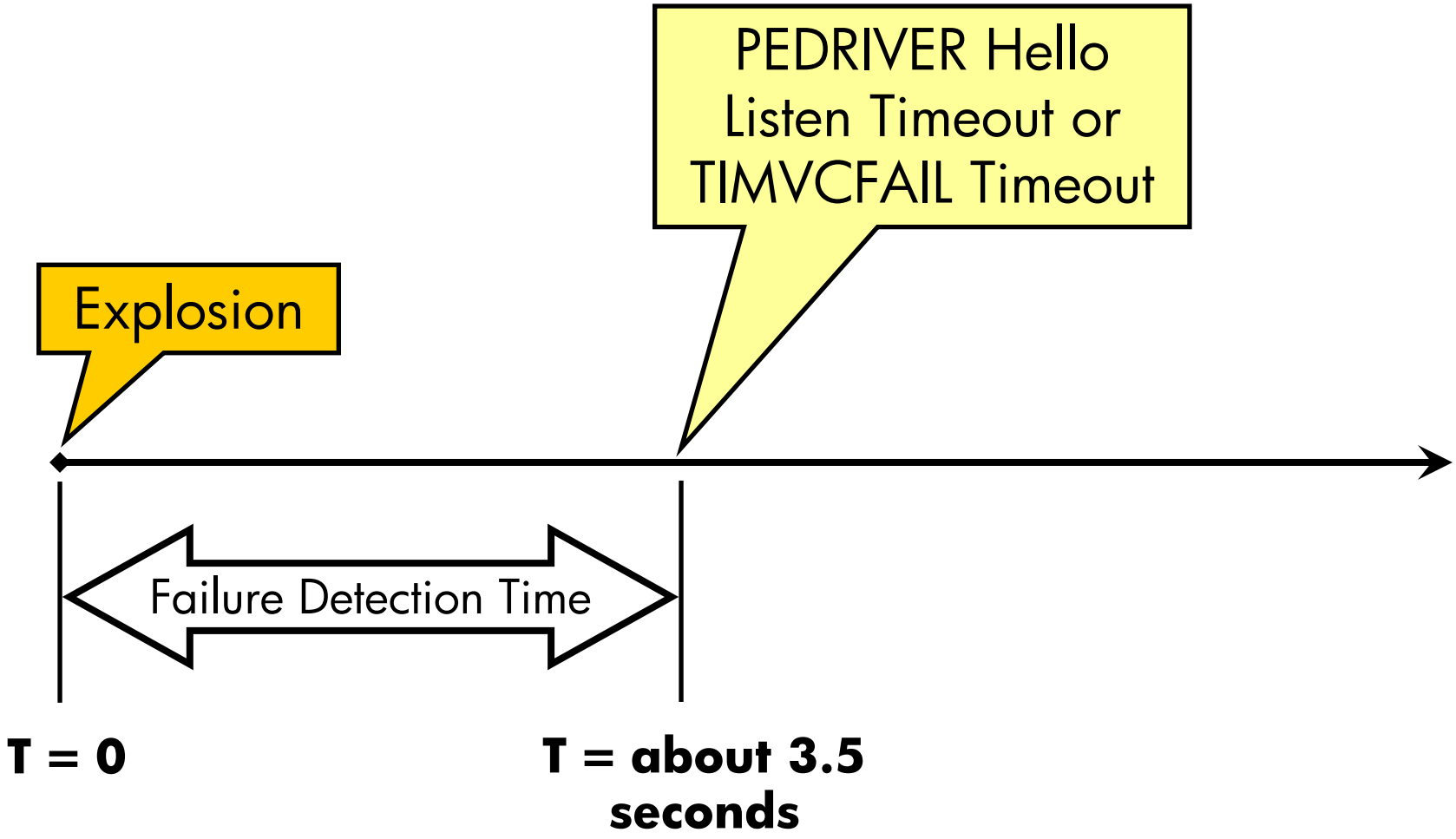


OpenVMS Disaster-Proof configuration & application

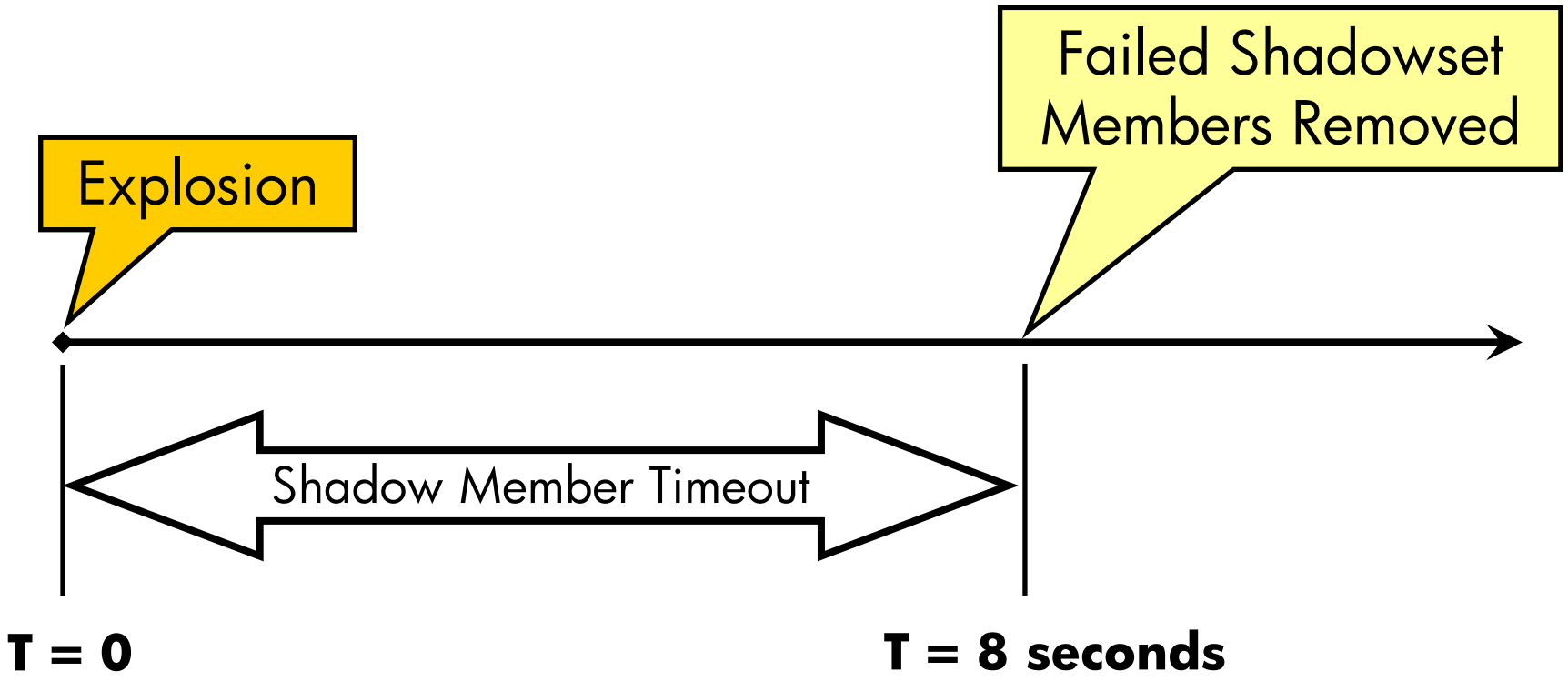


The longest outstanding request for an I/O during the DP demo was 13.7 seconds.

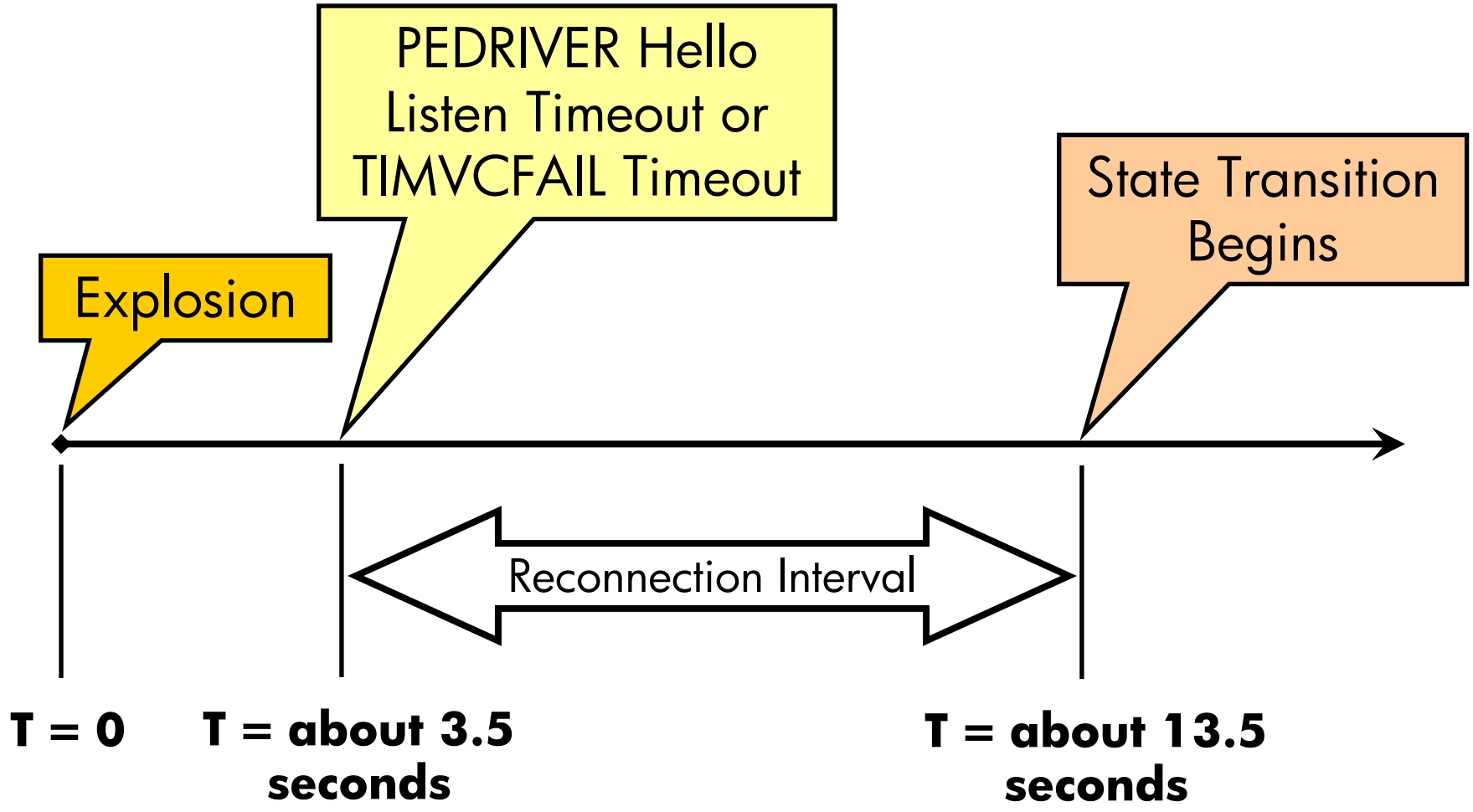
Disaster Proof Demo Timeline



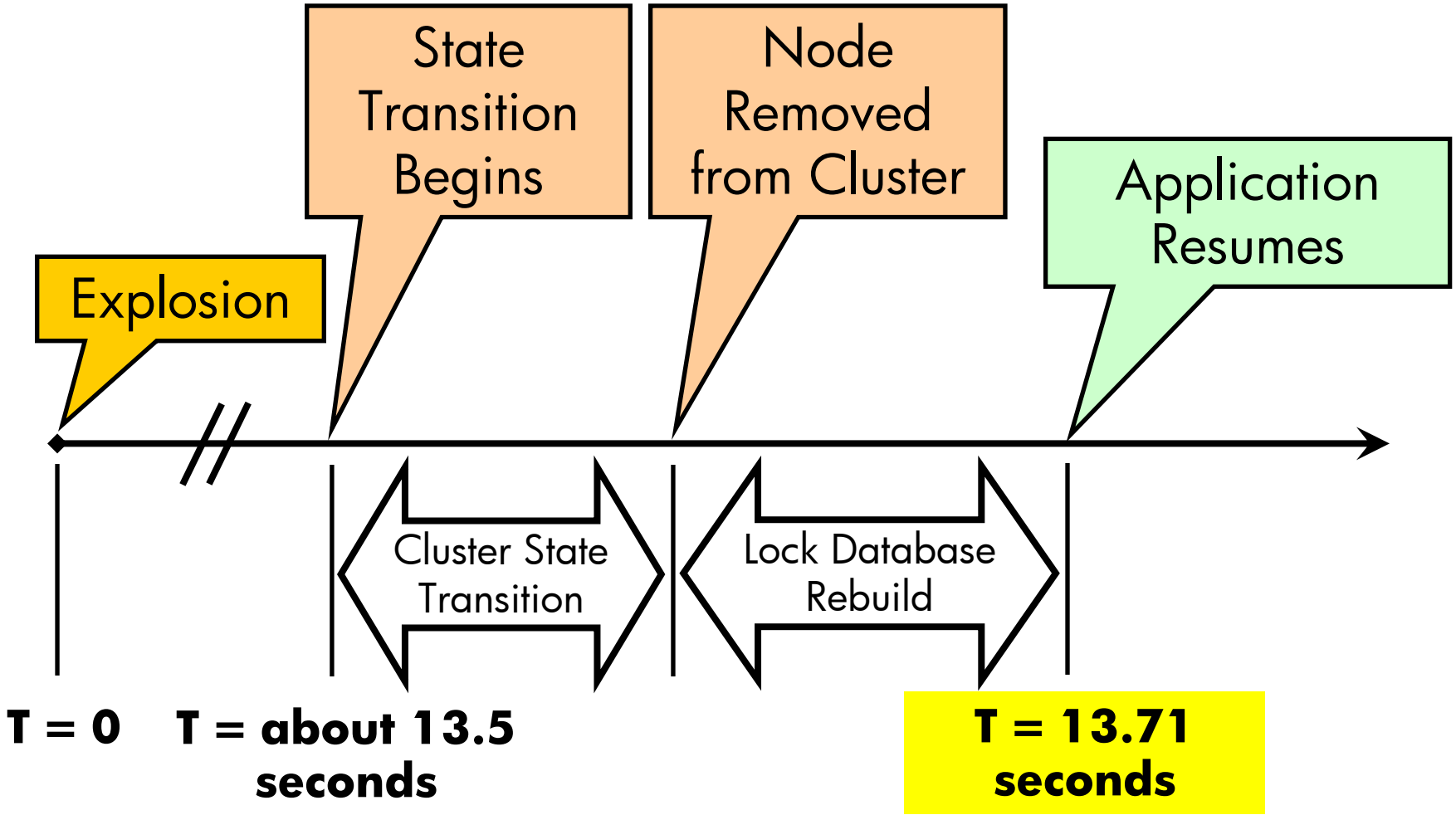
Disaster Proof Demo Timeline



Disaster Proof Demo Timeline



Disaster Proof Demo Timeline



Disaster Proof Demo Timeline

- Time = 0: Explosion occurs
- Time around 3.5 seconds: Node failure detected, via either PEDRIVER Hello Listen Timeout or TIMVCFAIL mechanism. VC closed; Reconnection Interval starts.
- Time = 8 seconds: Shadow Member Timeout expires; shadowset members removed.
- Time around13.5 seconds: Reconnection Interval expires; State Transition begins.
- Time = 13.71 seconds: Recovery complete; Application processing resumes.

OpenVMS System Parameter Settings for the Disaster Proof Demonstration

- SHADOW_MBR_TMO lowered from default of 120 down to 8 seconds
- RECNXINTERVAL lowered from default of 20 down to 10 seconds
- TIMVCFAIL lowered from default of 1600 to 400 (4 seconds, in 10-millisecond clock units) to detect node failure in 4 seconds, worst-case, (detecting failure at the SYSAP level)
- LAN_FLAGS bit 12 set to enable Fast LAN Transmit Timeout (give up on a failed packet transmit in 1.25 seconds, worst case, instead of an order of magnitude more in some cases)
- PE4 set to hexadecimal 0703 (Hello transmit interval of 0.7 seconds, nominal; Listen Timeout of 3 seconds), to detect node failure in 3-4 seconds at the PEDRIVER level

References

- Good Success Stories for OpenVMS DT clusters

http://h71000.www7.hp.com/success-stories.html

- Good presentations on DT:

http://www2.openvms.org/kparris/

- Spreadsheet to calculate latency due to speed of light over a distance
  - http://www2.openvms.org/kparris/Latency_due_to_Speed_of_Light.xls
- Disaster Tolerant Management Services from HP Services

http://h20219.www2.hp.com/services/cache/10597-0-0-225-121.html

- Disaster proof Video

http://hp.com/go/disasterproof/

Questions?