



GET CONNECTED

PEOPLE. TECHNOLOGY. RESULTS.

OpenVMS Disaster Tolerance Update

Keith Parris

Systems/Software Engineer, HP

June 17, 2009 Session 3049



Overview

- Disaster Risks
- Trends
- Case Studies

Disaster Risks



Mortal Hazard Risks in the United States

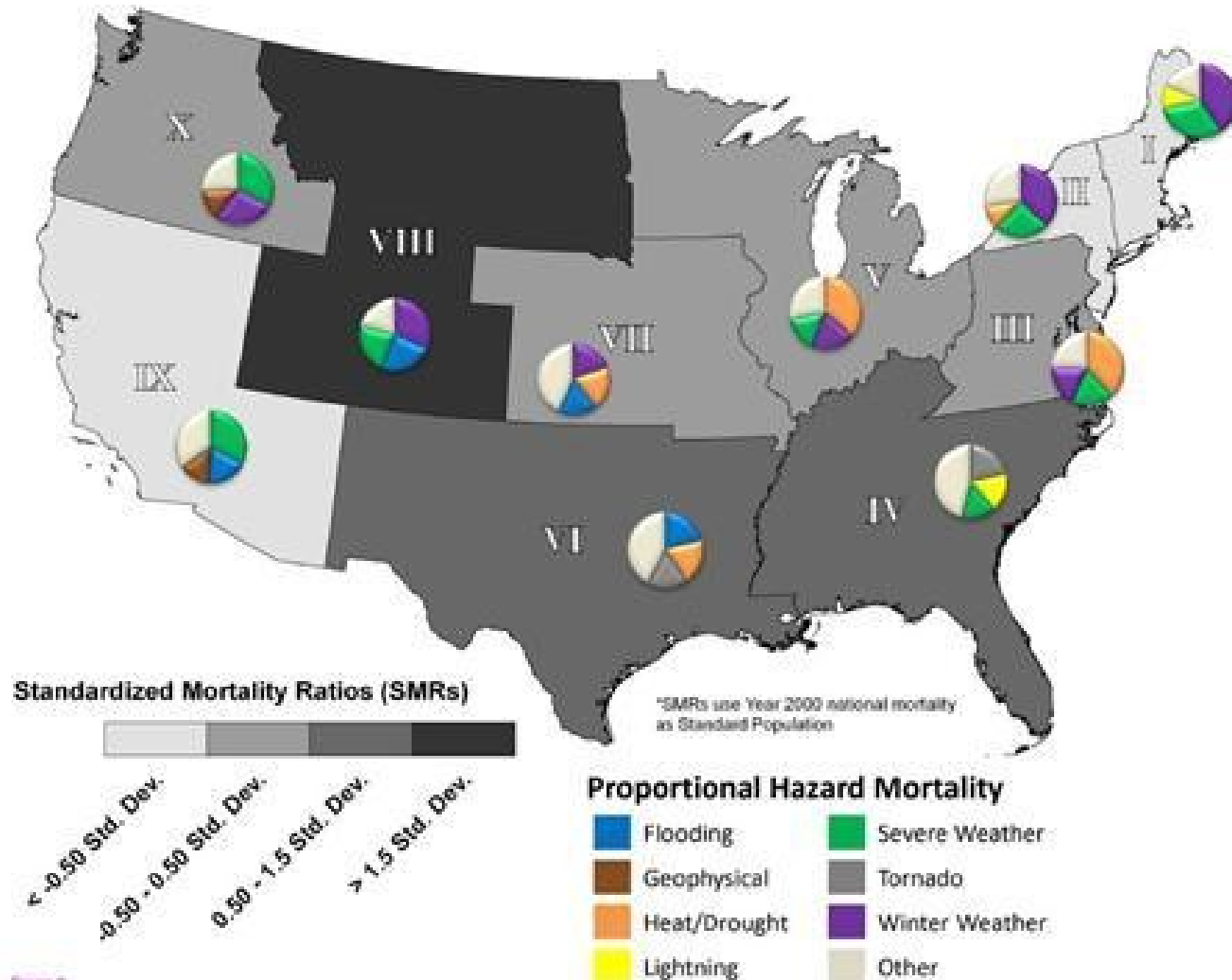


Figure 2

"Death map" shows heat a big hazard to Americans
Reuters, December 17, 2008

<http://www.reuters.com/article/healthNews/idUSTRE4BG01H20081217?feedType=RSS&feedName=healthNews>



“Every July the country's leading disaster scientists and emergency planners gather in Boulder, Colorado. ... They placed a ballot box next to the water pitchers and asked everyone to vote: What will be the next mega-disaster? A tsunami, an earthquake, a pandemic flu? And where will it strike?”

“Why We Don't Prepare for Disaster”

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



“The winner, with 32% of the votes, was once again a hurricane. After all, eight of the 10 costliest disasters in U.S. history have been hurricanes. This time, most of the hurricane voters predicted that the storm would devastate the East Coast, including New York City.”

“Why We Don't Prepare for Disaster”

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



“Here's one thing we know: a serious hurricane is due to strike New York City, just as one did in 1821 and 1938. Experts predict that such a storm would swamp lower Manhattan, Brooklyn and Jersey City, N.J., force the evacuation of more than 3 million people and cost more than twice as much as Katrina. An insurance-industry risk assessment ranked New York City as No. 2 on a list of the worst places for a hurricane to strike; Miami came in first.”

“Why We Don't Prepare for Disaster”

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



"If it seems like disasters are getting more common, it's because they are... Floods and storms have led to most of the excess damage. The number of flood and storm disasters has gone up 7.4% every year in recent decades, according to the Centre for Research on the Epidemiology of Disasters."

"Why Disasters Are Getting Worse

By Amanda Ripley

TIME Magazine, September 3, 2008

<http://www.time.com/time/nation/article/0,8599,1838400,00.html>



"We are getting more vulnerable to weather mostly because of where we live. ... In recent decades, people around the world have moved en masse to big cities near water. ... So the same-intensity hurricane today wreaks all sorts of havoc that wouldn't have occurred had human beings not migrated."

"Why Disasters Are Getting Worse

By Amanda Ripley

TIME Magazine, September 3, 2008

<http://www.time.com/time/nation/article/0,8599,1838400,00.html>



“Modern Americans are particularly, mysteriously bad at protecting themselves from guaranteed threats. We know more than we ever did about the dangers we face. But it turns out that in times of crisis, our greatest enemy is rarely the storm, the quake or the surge itself. More often, it is ourselves.”

“Why We Don't Prepare for Disaster”

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



"About half of those surveyed said they had personally experienced a natural disaster or public emergency. But only 16% said they were "very well prepared" for the next one. Of the rest, about half explained their lack of preparedness by saying they don't live in a high-risk area. In fact, 91% of Americans live in places at a moderate-to-high risk of earthquakes, volcanoes, tornadoes, wildfires, hurricanes, flooding, high-wind damage or terrorism."

"Why We Don't Prepare for Disaster"

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



“‘There are four stages of denial,’ says Eric Holdeman, director of emergency management for Seattle's King County, which faces a significant earthquake threat. ‘One is, it won't happen. Two is, if it does happen, it won't happen to me. Three: if it does happen to me, it won't be that bad. And four: if it happens to me and it's bad, there's nothing I can do to stop it anyway.’”

“Why We Don't Prepare for Disaster”

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



“Historically, humans get serious about avoiding disasters only after one has just smacked them across the face.”

“Why We Don't Prepare for Disaster”

TIME Magazine, August 20, 2006

<http://www.time.com/time/magazine/article/0,9171,1229102,00.html>



Trends



Disaster-tolerant cluster trends

- Distance Trends:
 - Longer inter-site distances for better protection (or because the company 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 “sweet spot” for cluster interconnects
 - IP network focus; increasing pressure not to bridge LANs between sites



Disaster-tolerant cluster trends

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

Disaster-Tolerant Cluster Customer Needs and OpenVMS Cluster Features in Response

- Site outage which is temporary [OpenVMS 8.3 allows Mini-Merge bitmaps to be converted to Mini-Copy bitmaps for quick recovery from unscheduled site outage)
- IP routing as network sweet spot, not bridging [thus OpenVMS plans to support IP as a Cluster Interconnect]
- Desire to still have redundancy of storage after a site failure in a disaster-tolerant cluster [thus OpenVMS plans to support up to 6-member shadowsets compared with the current limit of 3-member]



Case Studies



Case Study: 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
 - Round-trip latency of about 350 milliseconds
 - Estimated circuit path length about 22,000 miles



Case Study: 3,000-mile Cluster

- Unidentified OpenVMS customer
- 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: 1,400-mile Shadowing

- Healthcare customer
- 1,400 mile site separation distance
 - 1,875 mile circuit length; 23-millisecond round-trip latency
- Remote Vaulting (not clustering) across distance
- Synchronous Replication using Volume Shadowing
- Independent OpenVMS Clusters at each site (not cross-site)
- SAN Extension over OC-3 links with Cisco MDS 9000 series FCIP boxes
 - Writes take 1 round trip (23 millisecond write latency)
- Did things right; tested first with inter-site distance simulated using latency of 30 milliseconds with network emulator box
- Caché database – seems to be tolerant of high write latency and doesn't let those block reads



Case Study: 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
 - Round-trip latency 13 milliseconds
 - Estimated circuit path length about 800 miles
- Month-end processing time is one of the most performance-critical tasks
- Tested in OpenVMS Customer Lab using D4
- Performance impact too high:
 - May do shorter-distance DT cluster to new site, then use CA (Asynchronous) to distant site for DR purposes



Case Study: 20-mile DT Cluster

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

Case Study: 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
 - With default Read_Cost values, customer was getting all the detriment of Volume Shadowing for writes, but none of the benefit for reads



Case Study: 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:
 - $0.4 + (\text{twice } 0.8 \text{ millisecond round-trip time}) = 2.0 \text{ milliseconds}$
 - Factor of 5X slower write latency
- 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:
 - $0.4 + (\text{once } 0.8 \text{ millisecond round-trip time}) = 1.2 \text{ milliseconds}$
 - Factor of 3X slower write latency instead of 5X



Case Study: 20-mile DT Cluster

Read Latencies

- Shadowset member disks contain identical data, so Shadowing can read from any member disk
- When selecting a shadowset member disk for a read, Volume Shadowing adds the local queue length to the Read_Cost value and selects the disk with the lowest total to send the read to.
- Default OpenVMS Read_Cost values:
 - Local Fibre Channel disks = 2
 - MSCP-served disks = 501
 - Difference of 499
- Queue length at local site would have to reach 499 before sending any reads across to the remote site



Case Study: 20-mile DT Cluster

Read Latencies

- Example: 50-mile circuit path length, 0.8 millisecond round-trip latency, average local read latency measured at 0.4 milliseconds
- Read latency to remote disks calculated as:
 - $0.4 + (\text{one } 0.8 \text{ millisecond round-trip time for MSCP-served reads}) = 1.2 \text{ milliseconds}$
 - 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



Case Study: 20-mile DT Cluster

- Workaround: 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

Case Study: 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



Questions?



Speaker Contact Info



E-mail:

Keith.Parris@hp.com

Website:

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

Produced in cooperation with:

