DELL EMC

PowerEdge Product Group

Direct from Development

PowerEdge FX2 Optimized Fault Zone Design

Technical Note by:
Todd Mottershead

SUMMARY

A “Fault Zone” is the area within an infrastructure where the failure of a hardware subsystem can negate software based High Availability schemes like clusters.

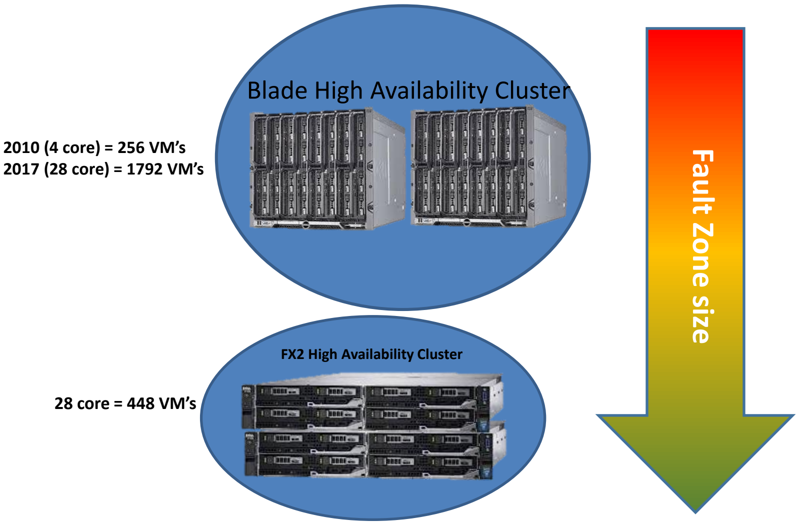
As CPU core counts increase, compute infrastructure scales to host an ever increasing number of virtual machines. This consolidation can result in a dramatic increase in the size of the Fault Zone.

FX2 Modular systems have been designed to reduce the size of the fault zone and thus, can significantly reduce the level of risk for customers.

Background

The complexity of managing large numbers of servers led to the development of blade systems where multiple servers could be enclosed and managed from a single chassis and cable/power consolidation could reduce the complexity of IT infrastructures. The challenge this approach created was that downtime incurred on the chassis affected all of the servers installed in it.

To address this, software High Availability (HA) from vendors like VMware and Microsoft evolved to become widely used by customers to protect their applications from unplanned downtime. When implemented in a blade environment, software vendors strongly recommend that “primary hosts” be distributed across multiple blade enclosures to ensure up-time in the event of an enclosure failure. This approach was effective in the past but as core-counts increase, customers face a tradeoff between consolidation and high availability. As the graphic below illustrates, a typical enclosure with 16 blades could support up to 256 virtual machines using 4 core CPU’s but with 28 core CPU’s, this capacity increases to over 1700 virtual machines*.


Blade High Availability Cluster
2010 (4 core) = 256 VM’s
2017 (28 core) = 1792 VM’s
FX2 High Availability Cluster
28 core = 448 VM’s
Fault Zone size

For some customers, their virtual machine requirements simply aren’t scaling at a rate that demands this level of capacity. For these customers, increased core counts result in excess capacity since they still require 2 enclosures but can utilize only a portion of each. For the many customers who can utilize this level of scale, the potential for the loss of over 1700 virtual machines due to a single hardware event (or chassis maintenance) represents an unacceptable level of risk even when utilizing Software based HA protection.

* based on 2 VM’s per core

Designing a Fault Zone to reduce Risk

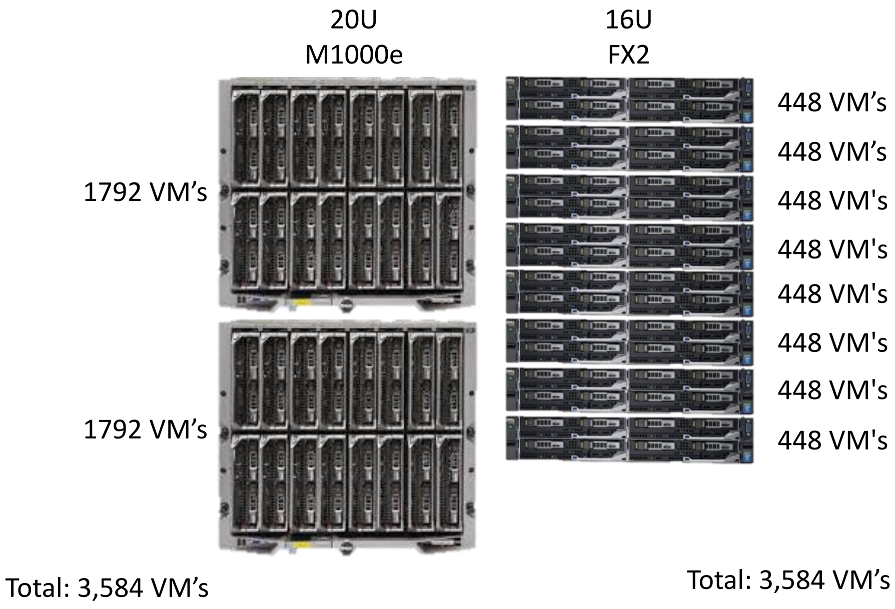
The two key concepts that define risk are: “Probability of Failure” and “Severity of Failure”. In simple terms, “how likely is the system to fail?” and “how much is it going to hurt if it fails?”. When designing a Fault Zone optimized architecture, both of these elements are important. From an engineering perspective, the “Probability of Failure” is a function of the number of components in the design and the annual failure rate of each of those components while “Severity of Failure” is a function of the amount of workload downtime associated with a failure. Reducing the “Probability of Failure” requires designing a system with fewer parts in each Fault Zone. Reducing “Severity of Failure” requires an architecture that spreads the risk by designing more, smaller Fault Zones.

FX2 Design elements that affect the Fault Zone - *Node Count*

Probability: Using the FC630/FC640 as examples, up to 4 nodes can be installed in a single FX2 Chassis. This represents a 75% reduction in the number of servers compared to a 16 node blade chassis which equates to fewer parts installed per chassis and thus a lower probability of failure for each fault zone.

Severity: Comparing a pair of 16 node blade chassis with a pair of 4 node FX2 chassis with both protected by vSphere High Availability, the FX2 solution would support up to 448 clustered virtual machines while the blade system would support up to 1792*. This equates to 75% fewer workloads affected by downtime on a chassis. Should the customer require additional virtual machines, they can add FX2 systems with confidence knowing that each pair represents an independent fault zone.

To illustrate, consider the graphic below. In this example, virtual machine capacity is the same however planned (or unplanned) downtime on an M1000e chassis would reduce capacity by 50% while downtime on the FX2 system would reduce capacity by only 12.5%. If the VM’s were provisioned for high availability across 2 chassis, downtime on an M1000e chassis would result in a complete loss of redundancy while downtime of an FX2 chassis would result in a 25% loss of redundancy. In practice, customers will typically overprovision to ensure enough capacity is available to maintain full redundancy but with an M1000e, this would require a third chassis (with 16 blades) and an additional 10U of space while the same could be accomplished with FX2 with the addition of only 2 enclosures (with 4 nodes each) in only 4U (with the total size equal to the space used by two M1000e enclosures).



FX2 Design elements that affect the Fault Zone - *Component reliability*

To reduce the probability of failure, all PowerEdge Servers are designed for the highest level of quality. The *Direct from Development* tech notes listed below apply to reducing component and manufacturing risk in FX2 and are available for download from Dell TechCenter, SalesEdge, and inside.dell.com. To download from TechCenter

1) Go to the **TechCenter Extras** site on TechCenter (dell.com/techcenter , then click on the TechCenter Extras tab
2) Click on "**White Papers and Media**"
3) Search for "**Direct from Development**"
    - PowerEdge Quality from Concept to EOL http://en.community.dell.com/techcenter/extras/m/white_papers/20443609
    - Reliability in Dell EMC PowerEdge Servers http://en.community.dell.com/techcenter/extras/m/white_papers/20444283
    - Understanding Our Unique Factory Process http://en.community.dell.com/techcenter/extras/m/white_papers/20443616
    - Why Choose Enterprise-class Drives http://en.community.dell.com/techcenter/extras/m/white_papers/20443820
    - Proactive High Availability with OMIVV http://en.community.dell.com/techcenter/extras/m/white_papers/20444536

Other Considerations in design

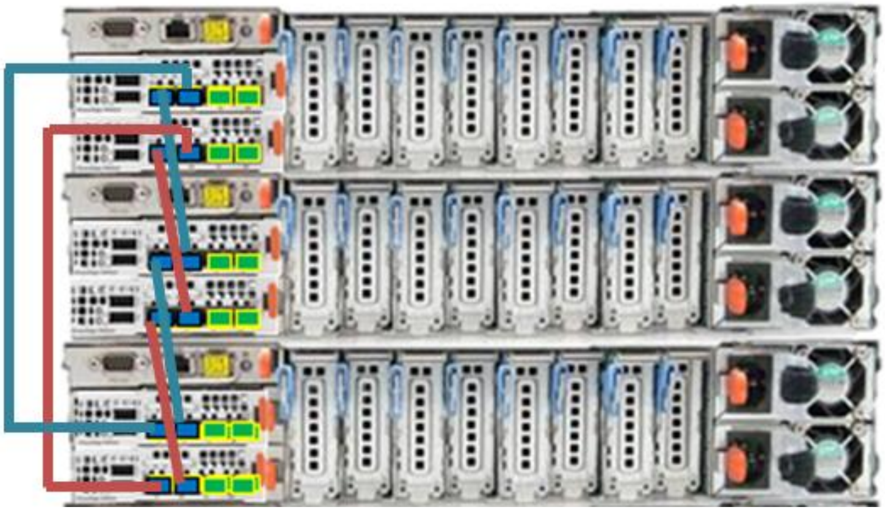
Datacenter space is expensive and recognizing that simply installing fewer servers in a blade chassis is too inefficient for many customers, Dell EMC Engineers also took the opportunity to reduce the size of the FX2 platform to 2U. This means that in the 10U space required by an M1000e with 16 nodes, customers can deploy up to twenty nodes across five FX2 chassis, and benefit from 25% more compute capacity while at the same time, significantly reducing fault zone severity.

Redundancy

To further reduce the level of risk, each FX2 chassis can be provisioned with redundant power supplies, redundant network paths (as described above), redundant management connections and additional redundancy can be obtained through PCIe cards for network or fiber channel connection. All of these capabilities work to reduce the probability of failure within the enclosure.

Networking

When reducing the size of the fault zone, the networking subsystem is critical for maintaining communications between enclosures both for cluster heartbeat functions as well as for establishing robust East/West network traffic flows for vMotion. FX2 systems benefit from the use of IO Modules which allow the customer to seamlessly establish uplinks to their top of rack switches, to segment redundant pathways for data and to link together, across up to 6 enclosures, for dedicated east/west traffic flows for vMotion and heartbeat. The graphic below highlights the cabling for 3 enclosures where the IO Modules are “stacked” to provide bi-directional data flow between all enclosures. This model allows the user to establish a performance zone for east/west traffic without affecting the fault zone while still providing 40Gb/s of uplink per enclosure to the top of rack switches.



Conclusion

Processor core counts are expected to continue to increase. While this trend provides significant benefits in terms of performance and scalability, it also creates risk. Consolidating workloads onto fewer and fewer servers reduces the probability of failure but dramatically increases the severity of failure. With these two conflicting trends in mind, Dell EMC Engineering teams worked to develop a new class of modular computing platform that could address this. By implementing a smaller fault zone, the PowerEdge FX2 design reduces the “severity of failure” compared to traditional architectures, while the built-in redundancy elements coupled with the world class Dell EMC quality processes address the “probability of failure” elements.

Whether a customer is looking to consolidate onto smaller systems or grow larger systems, the Fault Zone-optimized architecture of PowerEdge FX2 can help them reduce risk in their IT operations.