Understanding FX2 Power Management

The PowerEdge FX2 Shared Infrastructure Chassis implements the most accurate and highest performing power management scheme in Dell PowerEdge history.

Abstract

This paper provides a high-level description of the key aspects of the PowerEdge FX2 Shared Infrastructure Chassis Power Management scheme and emphasizes some of the key differences with previous PowerEdge modular chassis

November 2019

Dell EMC Technical White Paper

Revisions

Date	Description
Nov 2019	Initial release

Acknowledgements

This paper was produced by the following member of the Dell EMC Server engineering member:

Author

Kyle Cross — Technical Staff in the Dell EMC Server Solutions Group and the lead of the Modular Systems Engineering Team specializing in the design of blade and shared infrastructure power features

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Contents

Introduction

"Power Management" can refer to a lot of different features, but in this context, it means evaluating the power capacity available to the system from the Power Supply Units (PSUs) and ensuring with the highest performance efficiency possible that the system does not consume more than that total available power capacity.

Like all Dell PowerEdge products, power management of the PowerEdge FX2 Shared Infrastructure Chassis is divided into two very distinct functionalities: Power Budget Checks and Power Controls. In the following sections we will explain these functionalities in more detail.

Power Budgeting Protection Mechanisms in the PowerEdge Server Portfolio explains the reasoning behind Power Budget Checks and provides more detail on the types of Power Controls. This whitepaper will focus on the application of these concepts to the PowerEdge FX2 Shared Infrastructure Chassis.

1 Power Budget Checks

When compared to the previous PowerEdge M1000e and VRTX modular chassis, customers will notice some similarities and some differences in the PowerEdge FX2 Shared Infrastructure Chassis power management scheme. One aspect that all three chassis have in common is that each of them implement Power Budget Checks that ensure that as new components are given permission to power on they do not create a risk of exceeding the chassis power capacity. In each of the three chassis, the power capacity targets for the Power Budget Checks are based on the user’s selection of a Redundancy Policy (see Redundancy Policies). More conservative Redundancy Policies use lower power capacity targets, and depending on a customer’s configuration, may result in some chassis components being denied the ability to power on in order to ensure there is no risk of chassis shutdown. While these fundamentals remain the same, the “non-allocation” design of FX2 chassis is a significant difference from the M1000e and VRTX chassis which warrants further explanation.

The M1000e and VRTX Chassis use a "Power Allocation" scheme both for Power Budget Checks and as a Power Control scheme (see Power Control below). Before M-series PowerEdge Blades are turned on in an M1000e or VRTX chassis, the Dell Chassis Management Controller (CMC) performs a Power Budget Check to ensure there is sufficient available power capacity.

The PowerEdge FX2 Shared Infrastructure Chassis leverages the Power Budget Checks from the M1000e and VRTX chassis, and this is why it is sometimes referred to as an "allocation". Each of these three chassis perform initial Power Budget Checks to ensure that powering on a node will not violate the customer's requested Redundancy Policy setting. In the M1000e and VRTX chassis, this power on value is also used to “allocate” operational power amongst the installed blades and each blade is programmed to constrain itself to the allocation it is given. In the FX2 chassis, however, this Power Budget Check serves only to establish maximum and minimum bounds of operation for the installed sleds. The real time power consumption of sleds in the FX2 chassis is not constrained by this "power allocation” and is instead managed by real time chassis hardware controls (see Power Control below). This is distinctly different from the M1000e and VRTX chassis in which the "power allocation" instead defines a narrow range of operation that each blade is constrained to operate within (managed by the CMC, rather than chassis hardware).

In this way, the PowerEdge FX2 Shared Infrastructure Chassis substantially minimizes the role "power allocations" play in managing sled power consumption, and this is why FX2 is sometimes referred to as a "non-allocation" chassis. The benefits of this scheme will be explained in more detail in the Power Control section.

2 Power Control

Unlike the Power Budget Check, the Power Controls used in the PowerEdge FX2 Shared Infrastructure Chassis are fundamentally different from the M1000e and VRTX modular chassis.

The "Power Allocation" scheme used in M1000e and VRTX constrains each M-series blade to draw less than or equal to the power it has been allocated by the CMC. As power consumption on a given blade increases, the CMC will detect this condition and allocate more power to that blade if power is available. If power is not available, the blade will throttle its power consumption to operate within the range of power it has been allocated.

The FX2 chassis uses a revolutionary real time, hardware-based Power Control scheme that significantly improves on previous "allocation" based designs. Instead of limiting sleds to what they are allocated based on budgeting calculations, each sled is allowed to operate within a power limit that is continuously recalculated based on the real time power consumption of the overall FX2 chassis. When surplus power is available from the chassis PSUs, sleds are all allowed to operate up to maximum performance based on their workload needs. When power in the chassis becomes limited fast hardware, logic limits the power consumption of sleds to fit within chassis bounds. Unlike the M1000e and VRTX chassis there is no concept of "priority" in the FX2 chassis, and instead sled power consumption is limited in order of highest to lowest power consuming sled.

The new hardware-based Power Control scheme of the FX2 chassis allows it to pack in even denser configuration options than its bigger M1000e and VRTX cousins by optimizing the efficiency with which power is shared. "No allocations" means that power is never tied up in an allocation to a server node that isn't fully utilizing it. When power is available, sleds are able to increase power consumption and server performance freely. When workloads decrease and less power is needed, that power is instantly returned to the chassis power pool for use by other sleds. This hardware-based Power Control scheme is an essential part of what makes the incredible density of the PowerEdge FX2 Shared Infrastructure Chassis possible.

3 Redundancy Policies

Like other PowerEdge products, the PowerEdge FX2 Shared Infrastructure Chassis features user-configurable Redundancy Policy options. The user selection of Redundancy Policy affects the available Power Capacity used for Power Budget Checks and Power Control targets. More conservative Redundancy Policy settings may restrict server power on and set power throttling targets that limit server performance in order to ensure there is no risk of chassis shutdown in the event of an AC Grid or Power Supply failure.

The Dell Chassis Management Controller Version 1.3 for PowerEdge FX2/FX2s User's Guide provides a brief description of each of the four supported Redundancy Policies in the FX2 chassis. The four policies are:

- Grid Redundancy
- No Redundancy
- Redundancy Alerting Only
- Fault Tolerant Redundancy ***NEW***

In the following sections we'll address each of these modes as they relate to the Power Budget Checks and Power Controls described earlier in this document.

3.1 Grid Redundancy

3.1.1 Policy Budgeting

Grid or 1+1 Redundancy is a traditional redundancy mode which uses the power capacity limits of a single Power Supply Unit for Power Budget Checks. When the maximum potential power needs of installed chassis components exceed the capacity of a single Power Supply, the Chassis Management Controller (CMC) will deny power on to further chassis components. The Power Budget Checks for Grid Redundancy ensure that the PowerEdge FX2 Shared Infrastructure Chassis will remain operational in the event of maximum potential workload conditions at the time of an AC Grid or PSU Supply failure. Using the maximum potential is a conservative target that ensures continued operation across the wide range of potential customer workloads for a given configuration.

3.1.2 Policy Philosophy

In this way, Grid Redundancy is a conservative redundancy policy that ensures that the PowerEdge FX2 Shared Infrastructure Chassis and all installed components will remain operational with no risk of shutdown in the event of an AC Grid or Power Supply failure even when all installed components are simultaneously running at their worst-case power consumption.

3.1.3 Policy Control

As with all FX2 Redundancy Policies, while the two Power Supplies remain healthy, load is shared evenly between them and the capacity of both Power Supplies is made available for use. In the event of an AC Grid or Power Supply failure, Power Controls will rapidly engage to restrict the power consumption of the chassis and ensure that consumption is restricted to what a single Power Supply can support. For a fully loaded chassis running at maximum potential power this can result in some observed performance reduction as the chassis Power Control limits are enforced. In practice, customer workloads are often not at the maximum potential power and so practical performance reduction during an AC Grid or Power Supply failure is often minor or even completely unnoticeable.

3.1.4 Power On Behavior after Fault

In the event of an AC Grid or Power Supply failure, new chassis components will be allowed to power on provided that the maximum potential power of the newly installed chassis components does not exceed the capacity of a single Power Supply when evaluated by the chassis Power Budget Checks. This means that, while customers will note a chassis "Critical" state due to the loss of redundancy, they will observe no difference in which chassis components are allowed to power on (both before and after a redundancy fault). This is because in both cases, the chassis Power Budget Checks use the capacity of only a single Power Supply. This is a key difference from the more aggressive PowerEdge FX2 Shared Infrastructure Chassis Redundancy Policies.

3.1.5 Logging Behavior

As with all FX2 Redundancy Policies, when a Power Supply Unit fails, a log message will be generated. For the Grid Redundancy policy, a log message will also be recorded to note a "Loss of Redundancy". This message indicates that the system is continuing to operate in a Non-Redundant state, and action is necessary to either restore power to a failed AC Grid or replace a failed Power Supply Unit. Details in log messages make it possible to distinguish between these two cases. Finally, in the event that power on of a chassis component is denied due to a Power Budget Check, the denial will be logged both in CMC logs and iDRAC logs (in the case of compute sleds).

3.2 No Redundancy

3.2.1 Policy Budgeting

No Redundancy or 2+0 is the other traditional redundancy mode which uses the power capacity limits of both installed Power Supply Units for Power Budget Checks. The maximum potential power needs of installed chassis components are still used, but for this mode they are compared against the capacity of both installed Power Supply Units. For configurations where the summed Power Supply Unit capacity exceeds the overall power delivery capability of the chassis infrastructure, a "Chassis Infrastructure Limit" is also applied. This limit does not affect the Grid Redundancy policy selection, because single PSU capacities are not sufficient to exceed the Chassis Infrastructure Limit.

3.2.2 Policy Philosophy

The No Redundancy is an aggressive redundancy policy that prioritizes maximum performance over resiliency to AC Grid or Power Supply failures. In the event of such a failure, the chassis will still attempt to engage Power Controls to throttle and remain operational, but if Power Supply rated limits are exceeded the entire chassis will shut down to avoid potential hardware damage.

3.2.3 Policy Control

As with all FX2 Redundancy Policies, while the two Power Supplies remain healthy, load is shared evenly between them and the capacity of both Power Supplies is made available for use. When the No Redundancy policy is selected, in the event of an AC Grid or Power Supply failure, Power Controls will rapidly engage to restrict the power consumption of the chassis but may not be able to sufficiently constrain power to prevent a chassis shutdown. If shutdown can be avoided, extreme throttling may still be observed as the chassis Power Control limits are enforced. In practice, customer workloads will not always experience extreme throttling at the time off capacity loss, but the system remains in an exposed state since future load spikes may create power consumption levels which cannot be controlled which could lead to a chassis shutdown.

3.2.4 Power On Behavior after Fault

When the No Redundancy policy is selected, in the event of an AC Grid or Power Supply failure, new chassis components will be allowed to power on only if the maximum potential power of the newly installed chassis components does not exceed the capacity of a single Power Supply when evaluated by the chassis Power Budget Checks. This means that, customers may notice a difference in which chassis components are allowed to power on before and after an AC Grid or Power Supply failure. This is because after the capacity loss event the chassis Power Budget Checks use the capacity of only a single Power Supply (since that is all that is currently available in this case) and block the power on of components which do not fit into a single Power Supply capacity.

3.2.5 Logging Behavior

As with all FX2 Redundancy Policies, when a Power Supply Unit fails, a log message will be generated. For the No Redundancy policy, no separate redundancy message is logged. In the event that power on of a chassis component is denied due to a Power Budget Check, the denial will be logged both in CMC logs and iDRAC logs (in the case of compute sleds).

3.3 Redundancy Alerting Only

3.3.1 Policy Budgeting

The Redundancy Alerting Only policy is a hybrid redundancy mode which allows full use of the installed Power Supply Capacity for powering on chassis components but warns the user in cases where power exceeds limits which put the chassis at risk of shutdown in the event of redundancy loss. The Redundancy Alerting Only policy uses the power capacity limits of both installed Power Supply Units for Power Budget Checks. The maximum potential power needs of installed chassis components are still used, but for this mode they are compared against the capacity of both installed Power Supply Units. For configurations where the summed Power Supply Unit capacity exceeds the overall power delivery capability of the chassis infrastructure, a "Chassis Infrastructure Limit" is also applied. This limit does not affect the Grid Redundancy policy selection, because single PSU capacities are not sufficient to exceed the Chassis Infrastructure Limit.

3.3.2 Policy Philosophy

The Redundancy Alerting Only policy is a balanced redundancy policy that prioritizes maximum performance but complements that performance with warnings when AC Grid or Power Supply failures present a risk of shutdown. Warning messages are designed to give the user the opportunity to take action to reduce chassis load. In the event of an AC Grid Loss or Power Supply failure, the chassis will still attempt to engage Power Controls to throttle and remain operational, but if Power Supply rated limits are exceeded the entire chassis will shut down to avoid potential hardware damage. As a result, it is important to take action immediately on an FX2 system configured with Redundancy Alerting Only if redundancy loss warning messages occur.

3.3.3 Policy Control

As with all FX2 Redundancy Policies, while the two Power Supplies remain healthy, load is shared evenly between them and the capacity of both Power Supplies is made available for use. When the Redundancy Alerting Only policy is selected, in the event of an AC Grid or Power Supply failure, Power Controls will rapidly engage to restrict the power consumption of the chassis but may not be able to sufficiently constrain power to prevent a chassis shutdown. If shutdown can be avoided, extreme throttling may still be observed as the chassis Power Control limits are enforced. In practice, customer workloads will not always experience extreme throttling at the time off capacity loss, but the system remains in an exposed state since future load spikes may create power consumption levels which cannot be controlled which could lead to a chassis shutdown.

3.3.4 Power On Behavior after Fault

When the Redundancy Alerting Only policy is selected, in the event of an AC Grid or Power Supply failure, new chassis components will be allowed to power on only if the maximum potential power of the newly installed chassis components does not exceed the capacity of a single Power Supply when evaluated by the chassis Power Budget Checks. This means that, customers may notice a difference in which chassis components are allowed to power on before and after an AC Grid or Power Supply failure. This is because after the capacity loss event the chassis Power Budget Checks use the capacity of only a single Power Supply (since that is all that is currently available in this case) and block the power on of components which do not fit into a single Power Supply capacity.

3.3.5 Logging Behavior

As with all FX2 Redundancy Policies, when a Power Supply Unit fails, a log message will be generated. For the Redundancy Alerting Only policy, special additional "Redundancy Lost" alert messages will also be logged by the CMC. A message will be logged in the event that the CMC Power Budget Check determines that installed chassis components cannot be guaranteed to throttle down within the capacity of a single PSU, creating a future risk of chassis shutdown if an AC Grid or Power Supply fails. In addition, a message will be logged in the event that real time chassis power consumption exceeds the transient capacity of a single Power Supply which indicates that current chassis load levels could cause a shutdown if an AC Grid or Power Supply fails. Finally, in the event that power on of a chassis component is denied due to a Power Budget Check, the denial will be logged both in CMC logs and iDRAC logs (in the case of compute sleds).

3.4 Fault Tolerant Redundancy

3.4.1 Policy Budgeting

Fault Tolerant Redundancy is a hybrid redundancy mode which uses the power capacity limits of a single Power Supply Unit for Power Budget Checks, similar to Grid Redundancy, but enforces added performance limiting after redundancy is lost. It is only fully supported with FC640 modular sleds. Previous generation modular sleds will still work with Fault Tolerant Redundancy enabled, but they will treat it identically to Grid Redundancy.

When the maximum potential power needs of installed chassis components exceeds the capacity of a single Power Supply, the Chassis Management Controller (CMC) will deny power on to further chassis components. The Power Budget Checks for Fault Tolerant Redundancy ensure that the PowerEdge FX2 Shared Infrastructure Chassis will remain operational in the event of maximum potential workload conditions at the time of an AC Grid or PSU Supply failure. Using the maximum potential is a conservative target that ensures continued operation across the wide range of potential customer workloads for a given configuration.

3.4.2 Policy Philosophy

Similar to Grid Redundancy, Fault Tolerant Redundancy is a conservative redundancy policy that ensures that the PowerEdge FX2 Shared Infrastructure Chassis and all installed components will remain operational with no risk of shutdown in the event of an AC Grid or Power Supply failure even when all installed components are simultaneously running at their worst-case power consumption. New for Fault Tolerant Redundancy is a limit on peak performance that occurs when redundancy is lost. Fault Tolerant Redundancy is able to maintain the same conservative standards of redundancy as traditional Grid Redundancy by limiting peak power after redundancy is lost to levels which will fit within the surviving Power Supply.

3.4.3 Policy Control

As with all FX2 Redundancy Policies, while the two Power Supplies remain healthy, load is shared evenly between them and the capacity of both Power Supplies is made available for use. In the event of an AC Grid or Power Supply failure, Power Controls will rapidly engage to restrict the power consumption of the chassis and ensure that consumption is restricted to what a single Power Supply can support. In addition to controls used with all FX2 Redundancy Policies, Fault Tolerant Redundancy also implements additional performance limiting functionality which restricts the peak power after redundancy loss.

For a fully loaded chassis running at maximum potential power this can result in some observed performance reduction as the chassis Power Control limits are enforced. In practice, customer workloads are often not at the maximum potential power and so practical performance reduction during an AC Grid or Power Supply failure is often minor or even completely unnoticeable.

3.4.4 Power On Behavior after Fault

In the event of an AC Grid or Power Supply failure, new chassis components will be allowed to power on provided that the maximum potential power of the newly installed chassis components does not exceed the capacity of a single Power Supply when evaluated by the chassis Power Budget Checks. This means that, while customers will note a chassis "Critical" state due to the loss of redundancy, they will observe no difference in which chassis components are allowed to power on (both before and after a redundancy fault). This is because in both cases, the chassis Power Budget Checks use the capacity of only a single Power Supply. This is a key difference from the more aggressive PowerEdge FX2 Shared Infrastructure Chassis Redundancy Policies.

3.4.5 Logging Behavior

As with all FX2 Redundancy Policies, when a Power Supply Unit fails, a log message will be generated. For the Fault Tolerant Redundancy policy, a log message will also be recorded to note a "Loss of Redundancy". This message indicates that the system is continuing to operate in a Non-Redundant state, and action is necessary to either restore power to a failed AC Grid or replace a failed Power Supply Unit. Details in log messages make it possible to distinguish between these two cases. Finally, in the event that power on of a chassis component is denied due to a Power Budget Check, the denial will be logged both in CMC logs and iDRAC logs (in the case of compute sleds).