Übersicht

Die Fujitsu PRIMERGY Server sind bereits bei der Auslieferung ab Werk mit BIOS-Standardeinstellungen konfiguriert, die für die gängigsten Anwendungsszenarien ein optimales Verhältnis zwischen Performance und Energieeffizienz bieten. Dennoch gibt es Situationen, in denen es erforderlich sein kann, von den Standardeinstellungen abzuweichen und damit den Server je nach Anforderung für den maximalen Durchsatz, sprich höchste Performance bei möglichst geringer Latenz zu konfigurieren. Dies trifft insbesondere auf den High Performance Computing (HPC) Bereich und auf Anwendungen aus dem Finanzmarkt zu, wo es darum geht, Millionen von Transaktionen pro Sekunde und Daten ohne Verzögerung möglichst in Echtzeit zu verarbeiten. Neben diesem Szenario gibt es auch Umgebungen, in denen nicht die reine Performance die größte Rolle spielt, sondern die höchste Energieeffizienz. Dieses trifft z.B. auf Rechenzentrumsbetreiber zu, denen nur ein begrenztes Budget an elektrischer Leistung zur Verfügung steht. Dort wird versucht die Server so zu optimieren, dass diese möglichst viel Durchsatz, jedoch bei niedrigst möglicher elektrischer Leistungsaufnahme liefern. Für diese Szenarien werden in diesem White Paper Empfehlungen für optimale BIOS-Eistellungen gegeben.

PRIMERGY BIOS-Optionen

Dieses White Paper beinhaltet nur BIOS-Optionen, die für die Intel Xeon E5-2400/2600/4600 basierten PRIMERGY Server gelten. Das BIOS der PRIMERGY Server wird kontinuierlich weiterentwickelt. Deshalb ist es wichtig die jeweils neueste BIOS-Version zu verwenden, um alle hier aufgeführten BIOS-Optionen verfügbar zu haben. Die aktuelle BIOS-Version der PRIMERGY Server kann im Internet unter http://www.fujitsu.com/fts/support heruntergeladen werden.

Empfehlungen für Performance, geringe Antwortzeit und Energieeffizienz

In der folgenden Tabelle sind Empfehlungen für BIOS-Optionen aufgeführt, die den Server entweder für beste Performance, niedrige Antwortzeit oder für höchste Energieeffizienz optimieren. Um die BIOS-Optionen zu ändern, muss zunächst während des Selbsttests des Systems (Power On Self Test = POST) das BIOS-Setup aufgerufen werden. Weitere Informationen dazu sind im Handbuch des Servers zu finden.

Bevor Änderungen der in der folgenden Tabelle gelisteten BIOS-Optionen vorgenommen werden, empfiehlt es sich die Fußnoten und die anschließende Beschreibung der BIOS-Optionen zu beachten.

BIOS-Setup-Menü	BIOS-Option	Einstellungen [1]	Performance	Low-Latency	Energieeffizienz
Advanced > PCI Subsystem Settings	ASPM Support	**Disabled** Auto Limit to L0s	Disabled	Disabled	Auto
Advanced > PCI Subsystem Settings	DMI Control	**GEN 2** GEN 1	GEN 2	GEN 2	GEN 1 [2]
Advanced > CPU Configuration	Hyper-Threading	Disabled **Enabled**	Enabled	Disabled [3]	Enabled
Advanced > CPU Configuration	[Hardware] [Adjacent Sector] [DCU Streamer] [DCU Ip] Prefetcher	Disabled **Enabled**	Enabled	Enabled	Disabled
Advanced > CPU Configuration	Intel Virtualization Technology	Disabled **Enabled**	Disabled [4]	Disabled	Disabled
Advanced > CPU Configuration	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom
Advanced > CPU Configuration	Turbo Mode	Disabled **Enabled**	Enabled	Enabled	Enabled
Advanced > CPU Configuration	Energy Performance [5]	Performance **Balanced Performance** Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient
Advanced > CPU Configuration	CPU C3 Report [5]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C6 Report [5]	Disabled **Enabled**	Disabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C7 Report [5]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	Package C-State limit [5]	C0 C2 C6 C7 **No Limit**	C0	C0	No Limit
Advanced > CPU Configuration	QPI Link Frequency Select	**Auto** 6.4 GT/s 7.2 GT/s 8.0 GT/s	Auto	Auto	6.4 GT/s
Advanced > CPU Configuration	QPI Link Power Management	Disabled **Enabled**	Disabled	Disabled	Enabled
Advanced > CPU Configuration	Frequency Floor Override	**Disabled** Enabled	Disabled [6]	Enabled	Disabled
Advanced > Memory Configuration	DDR Performance	Low-Voltage optimized Energy optimized **Performance optimized**	Performance optimized	Performance optimized	Low-Voltage optimized
Advanced > Memory Configuration	Patrol Scrub	Disabled **Enabled**	Disabled	Disabled	Disabled
Advanced > USB Configuration	Onboard USB Controllers	Disabled **Enabled**	Enabled	Enabled	Disabled [7]

[1] Die fett gedruckte Einstellung ist der Standardwert.
[2] GEN 1 wird empfohlen bei niedriger Chipsatz I/O Auslastung; anderenfalls ist GEN 2 einzustellen.
[3] Wenn nicht alle Threads der CPU genutzt werden, kann das Ausschalten von Hyper-Threading die Latenz verbessern.
[4] Wenn Virtualisierung nicht genutzt wird, sollte diese Option auf „Disabled“ gesetzt werden.
[5] Diese Option ist nur sichtbar, wenn die Einstellung für „Power Technology“ auf „Custom“ geändert wird.
[6] Das Einschalten dieser Option kann bei Anwendungen, die nicht alle CPUs des Systems auslasten, von Vorteil sein.
[7] Das Ausschalten dieser Option verhindert die Nutzung von internen oder externen USB Geräten.

Beschreibung der BIOS-Optionen

ASPM Support

ASPM steht für „Active State Power Management“ und ermöglicht es, die PCIe-Links zu den PCIe-Geräten in unterschiedliche Stromsparzustände zu schicken um dadurch die Leistungsaufnahme zu reduzieren. Mit der Einstellung „Auto“ wählt das System je nach Aktivität des PCIe-Links den energieeffizientesten Stromsparzustand. Das Wechseln der Stromsparzustände bzw. das Aufwachen aus den unterschiedlichen Stromsparzuständen erhöht jedoch die Latenz. Für volle I/O-Performance der PCIe-Geräte sollte deshalb die Einstellung „Disabled“ gewählt werden.

DMI Control

DMI steht für „Digital Media Interface“ und stellt die Verbindung zwischen dem Intel Prozessor und dem Chipsatz dar. Dieser Link kann mit unterschiedlichen Geschwindigkeiten betrieben werden. Der Chipsatz stellt unter anderem die Kommunikation zu den onboard LAN-Controllern, USB-Controllern, onboard SAS/SATA-Controllern und gegebenenfalls auch zu PCIe Slots, etc. zur Verfügung. Für Umgebungen, in denen die vom Chipsatz bereitgestellte Kommunikation nur wenig genutzt wird, kann die Geschwindigkeit des DMI-Links von „GEN 2“ auf „GEN 1“ heruntergesetzt werden, um damit die Leistungsaufnahme zu reduzieren.

Hyper-Threading

Grundsätzlich empfiehlt Fujitsu „Hyper-Threading“ einzuschalten („Enabled“). Dennoch kann es für Anwendungen, die speziell auf möglichst geringe Antwortzeiten Wert legen, wie es z.B. bei Trading Software aus dem Finanzmarkt oder bei HPC-Anwendungen der Fall ist, sinnvoll sein Hyper-Threading auszuschalten („Disabled“). Anwender aus diesen Bereichen sind meist weniger an dem maximalen Systemdurchsatz interessiert, der durch die zusätzlichen Threads zur Verfügung gestellt wird, als vielmehr an der Performance und Stabilität eines einzelnen Threads. In den Fällen, in denen die zusätzlichen Threads nicht genutzt werden und die Antwortzeit minimiert werden soll, sollte Hyper-Threading im BIOS ausgeschaltet („Disabled“) werden.

Prefetcher

Das BIOS der PRIMERGY Server enthält mehrere Prefetcher-Optionen. Dazu gehören „Hardware Prefetcher“, „Adjacent Cache Line Prefetch“, „DCU Streamer Prefetcher“ und „DCU Ip Prefetcher”. Die Prefetcher sind Funktionen des Prozessors, die es erlauben Daten nach bestimmten Mustern vorab aus dem Hauptspeicher in den L1 oder L2-Cache des Prozessors zu laden. Das Einschalten („Enabled”) der Prefetcher sorgt in der Regel für eine höhere Cache-Trefferrate und steigert somit die Gesamtperformance des Systems. Eine Ausnahme bilden Anwendungsszenarien, bei denen der Hauptspeicher voll ausgelastet ist und die Speicheranbindung einen Performanceengpass darstellt. In solchen Fällen kann es von Vorteil sein, die Prefetcher Option auf „Disabled“ zu setzen, um so die Bandbreite, die sonst für das Prefetching aufgebraucht wird, zusätzlich nutzen zu können. Darüber hinaus kann durch das Ausschalten der Prefetcher auch die Leistungsaufnahme des Servers reduziert werden. Bevor die Prefetcher Optionen auf Produktivsystemen geändert werden, sollten in einer Testumgebung zunächst die Auswirkungen der einzelnen Einstellungen für das jeweilige Anwendungsszenario untersucht werden.

Details zu den einzelnen Prefetchers:

Hardware Prefetcher

Dieser Prefetcher sucht nach Datenströmen in der Annahme, dass wenn die Daten an Adresse A und A+1 angefordert werden, vermutlich auch die Daten an Adresse A+2 benötigt werden. Diese werden dann vorab in den L2-Cache aus dem Hauptspeicher geladen.

Adjacent Cache Line Prefetch

Dieser Prefetcher holt immer Cache Line Paare (128 Byte) aus dem Hauptspeicher, vorausgesetzt die Daten sind nicht bereits im Cache enthalten. Wenn dieser Prefetcher ausgeschaltet wird, wird immer nur eine Cache Line (64 Byte) geholt, welche die Daten beinhaltet, die der Prozessor gerade braucht.

DCU Streamer Prefetcher

Bei diesem Prefetcher handelt es sich um einen L1-Cache Daten Prefetcher, der erkennt, wenn innerhalb einer bestimmten Zeit mehrfach Daten aus einer Cache Line angefordert werden, um dann, in der Annahme die nächste Cache Line wird ebenfalls benötigt, diese dann vorab aus dem L2-Cache oder dem Hauptspeicher in den L1-Cache zu laden.

DCU Ip Prefetcher

Dieser L1-Cache Prefetcher sucht nach vorangegangenen sequentiellen Zugriffen und versucht, basierend darauf, die als nächstes zu erwartenden Daten zu ermitteln und gegebenenfalls vorab aus dem L2-Cache oder dem Hauptspeicher in den L1-Cache zu laden.

Intel Virtualization Technology

Diese BIOS-Option schaltet zusätzliche Virtualisierungsfunktionen der CPU ein („Enabled“) oder aus („Disabled“). Wird der Server nicht für Virtualisierung genutzt, sollte diese Option auf „Disabled“ gesetzt werden. Dies kann zu einer Energieeinsparung führen.

Power Technology

Die BIOS-Option „Power Technology“ ist eine Obermenge an unterschiedlichen BIOS-Optionen, die die Performance und die Power Management Funktionen der Prozessoren steuern. Die Standardeinstellung „Energy Efficient“ stellt bereits eine gute Balance zwischen elektrischer Leistungsaufnahme und Performance ein. Um die dazugehörigen Optionen einzublenden und einzeln einzustellen, kann die Einstellung „Custom“ gewählt werden um weitere Einstellungen für die BIOS-Optionen „Energy Performance“, „CPU C3/C6/C7 Report“ und „Package C-State limit“ vornehmen zu können. Neben diesen Einstellungen gibt es noch weitere Optionen, die hier nicht aufgeführt sind, weil nicht empfohlen wird, von deren Standardwert abzuweichen. Die Einstellung „Disabled“ deaktiviert das Power Management der Prozessoren und beschränkt damit gleichzeitig die maximale Prozessorfrequenz, durch das Ausschalten der „Turbo Mode“ Option, auf die Nominalfrequenz.

Turbo Mode

Diese BIOS-Option schaltet die Intel Turbo Boost Technology Funktion des Prozessors ein („Enabled“) bzw. aus („Disabled“). Die Turbo Boost Technology Funktion erlaubt den Betrieb des Prozessors mit höheren Frequenzen als der Nominalfrequenz. Die maximal erreichbare Frequenz ist je nach Prozessor-Typ unterschiedlich und darüber hinaus abhängig von der Anzahl aktiver Cores, der Stromzufuhr, der elektrischen Leistungsaufnahme und der Temperatur des Prozessors. Neben diesen Randbedingungen für die Turbo Mode Performance spielt, insbesondere bei HPC-Anwendungen, auch die Qualität der Prozessoren eine Rolle.

Grundsätzlich empfiehlt Fujitsu die „Turbo Mode“ Option auf der Standardeinstellung „Enabled“ zu belassen, denn durch die höheren Frequenzen wird die Performance deutlich gesteigert. Da die höheren Frequenzen jedoch abhängig sind von Randbedingungen und nicht immer garantiert sind, kann es für Anwendungsszenarien, in denen eine konstante Performance gefordert ist, von Vorteil sein die Turbo Mode Option auszuschalten („Disabled“).

Energy Performance

Diese BIOS-Option steuert je nach Einstellung die interne „Power Control Unit“ der Intel Prozessoren und optimiert die Power Management Funktionen der Prozessoren zwischen Performance und Energieeffizienz. Mögliche Einstellungen sind „Performance“, „Balanced Performance“, „Balanced Energy“ und „Energy Efficient“. Einige Betriebssysteme überschreiben diese Einstellung, je nachdem wie die entsprechenden Energiesparoptionen konfiguriert sind.

CPU C3/C6/C7 Report

Mit diesen BIOS-Optionen wird dem Betriebssystem mitgeteilt, ob es die CPU C3, C6 oder C7 States nutzen kann oder nicht. Die CPU C-States sind Idle-Zustände, in denen der Core eines Prozessors, wenn er keinen Code auszuführen hat, in eine Art Schlafzustand versetzt wird. Dadurch wird der Stromverbrauch im Idle deutlich reduziert. Da das Aufwachen aus diesen Schlafzuständen die Latenz erhöht, wird empfohlen, für Anwendungen bei denen es auf maximale Performance bei möglichst geringer Antwortzeit ankommt, die Einstellung für die CPU C-States auf „Disabled“ zu setzen. Dabei gilt, je höher der C-State, desto länger die Aufwachzeit. Dabei sollte beachtet werden, dass wenn alle CPU C-States ausgeschaltet sind, nicht mehr die höchst mögliche Turbo Mode Frequenz erreicht werden kann. In diesem Fall würde unabhängig von der Anzahl aktiver Cores, die höchste Turbo Mode Frequenz auf die maximale Frequenz begrenzt werden, die möglich ist, wenn alle Cores aktiv sind. Diese ist je nach Prozessor-Typ in der Regel deutlich niedriger.

Mit der Einstellung „Disabled“ für die BIOS Option „CPU C3/C6/C7 Report“, wird seitens BIOS nur verhindert, dass der entsprechende CPU C-State per ACPI an das Betriebssystem übergeben wird, welches dann in der Regel nicht mehr in der Lage ist diesen zu nutzen. Eine Ausnahme bilden Betriebssysteme

(dazu gehört z.B. Red Hat Enterprise Linux 6.2), die nicht per ACPI, sondern per Treiber die möglichen CPU C-States ermitteln. In solchen Fällen muss die Nutzung der CPU C-States im Betriebssystem unterbunden werden (z.B. über Kernel Parameter).

Package C-State limit

Neben den CPU oder Core C-States gibt es auch sogenannte Package C-States, die es erlauben, nicht nur den einzelnen Core eines Prozessors, sondern den gesamten Prozessor-Chip in eine Art Schlafzustand zu versetzen. Die Leistungsaufnahme wird dadurch nochmals weiter reduziert. Die „Aufwachzeit“, die benötigt wird, um aus den tieferen Package C-States in den aktiven C0 State zu wechseln, ist im Vergleich zu den CPU oder Core C-States noch größer. Wird im BIOS die Einstellung „C0“ eingestellt, dann bleibt der Prozessor-Chip immer aktiv.

QPI Link Frequency Select

Mit dieser BIOS-Option ist es möglich, die Geschwindigkeit des Interconnects (QPI) zwischen den CPUs in einem System zu reduzieren um damit Strom einzusparen. Dies macht insbesondere dann Sinn, wenn die verfügbare Bandbreite nicht notwendig ist. Ist jedoch maximale Performance und eine geringe Antwortzeit die Vorgabe, belässt man es bei der Einstellung „Auto“, die automatisch die höchste Geschwindigkeit einstellt. Je nachdem welche Bandbreite benötigt wird kann hier zwischen den Geschwindigkeiten „6.4 GT/s“, welche die größte Einsparung bringt, „7.2 GT/s“ und „8.0 GT/s“, welche die maximale Geschwindigkeit ist, ausgewählt werden.

QPI Link Power Management

Diese BIOS-Option ermöglicht es, die Stromsparfunktionen der QPI-Links zu aktivieren („Enabled“) bzw. zu deaktivieren („Disabled“). Ähnlich wie die CPUs Idle C-States haben, besteht auch bei den QPI-Links die Möglichkeit, diese in eine Art Schlafzustand zu versetzen, wenn ein oder beide Prozessoren idle sind. Die Einsparung durch das „QPI Link Power Management“ im Idle ist besonders groß. Genau wie bei den CPU C-States erhöht das Power Management jedoch die Latenz. Deshalb sollte es bei Anwendungen, bei denen es auf maximale Performance bei möglichst geringer Antwortzeit ankommt, abgeschaltet werden („Disabled“).

Frequency Floor Override

Das Einschalten dieser BIOS-Option sorgt dafür, dass der Prozessor immer mit seiner maximalen Nominalfrequenz arbeitet, auch dann, wenn er wenig zu tun hat. Dementsprechend ist der Stromverbrauch auch höher und deshalb sollte im Normalfall die Einstellung für diese Option immer „Disabled“ sein. Eine Ausnahme bilden Anwendungen, deren Threads nicht alle CPUs des Systems auslasten. In diesem Fall sind die Zugriffe auf die Remote-CPU z.B. für die Cache-Kohärenz und besonders Zugriffe auf den Remote-Speicher der anderen CPU, oder auf PCIe Geräte, die an der anderen CPU angebunden sind, deutlich langsamer. Um in diesem Fall die Latenz so gering wie möglich zu halten, kann die BIOS-Option „Frequency Floor Override“ auf „Enabled“ gesetzt werden, wenn die dadurch erhöhte elektrische Leistungsaufnahme in Kauf genommen wird.

DDR Performance

Diese BIOS-Option steuert die Geschwindigkeit und die Spannung, mit der die im System gesteckten Speichermodule betrieben werden. Dabei wird zwischen Performance und Energieverbrauch abgewägt. Die Einstellung „Performance optimized“ betreibt die DIMMs mit der Spannung 1.5 V und ermöglicht so die maximale Geschwindigkeit. Mit der Einstellung „Low-Voltage optimized“ werden die DIMMs, falls möglich, mit energiesparenden 1.35 V betrieben. Dieser Betrieb ist nur bei Speicherbestückungen mit einem oder zwei DIMMs pro Speicherkanal möglich und kann (Aufschluss darüber gibt das Memory Performance White Paper) die Speicherfrequenz begrenzen. Die Einstellung „Energy optimized“ begrenzt die Speicherfrequenz zusätzlich auf den minimalen Wert (800 MHz). Für optimale Energieeffizienz wird die Einstellung „Low-Voltage optimized“ empfohlen und für die maximale Speicher-Performance die Einstellung „Performance optimized“.

Neben den BIOS-Optionen für Speicher Performance spielen der verwendete Speicher-Typ und die optimale Bestückung der DIMMs eine weitaus größere Rolle. Eine ausführliche Beschreibung dazu sowie zu dem Thema NUMA findet man im Memory Performance White Paper (siehe Literaturliste am Ende des Dokuments).

Patrol Scrub

Diese BIOS-Option schaltet das sogenannte Memory Scrubbing ein („Enabled“) oder aus („Disabled“), welches unabhängig vom Betriebssystem im Hintergrund zyklisch auf den gesamten Hauptspeicher des Systems zugreift, um Speicherfehler präventiv aufzuspüren und zu korrigieren. Der Zeitpunkt dieses Speichertests ist nicht beeinflussbar und kann unter Umständen zu Performance-Einbußen führen. Das Ausschalten („Disabled“) der Patrol Scrub Option erhöht die Wahrscheinlichkeit, bei aktiven Zugriffen durch das Betriebssystem, auf Speicher Fehler zu stoßen. Solange diese Fehler korrigierbar sind, sorgt die ECC Technologie der Speichermodule dafür, dass das System stabil weiterläuft. Zu viele korrigierbare Speicher Fehler erhöhen jedoch das Risiko auf nicht-korrigierbare Fehler zu stoßen, die dann zum Systemstillstand führen.

Onboard USB Controllers

Der Chipsatz der PRIMERGY Server beinhaltet mehrere USB-Controller. Wenn auf den Einsatz von USB-Geräten komplett (dies beinhaltet auch Maus und Tastatur) verzichtet werden kann, sollte die Einstellung für diese BIOS-Option „Disabled“ sein. Dies spart Strom und erhöht die Sicherheit durch unerlaubten Zugriff Unbefugter. Unabhängig von der Einstellung bleiben die USB-Controller während des Systemstarts aktiv (die Deaktivierung erfolgt erst nach dem POST), so dass man auch bei der Einstellung „Disabled“ noch die Möglichkeit hat per USB Tastatur ins BIOS-Setup zu gelangen um die Einstellung wieder zu ändern.