Intel® C440GX+ Server Board ACPI White Paper

Revision 1.1
April 14, 1999

Revision History		
Date	**Rev**	**Modifications**
4/8/99	1.0	Initial release
4/14/98	1.1	Corrections to format

Information in this document is provided in connection with Intel products. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document. Except as provided in Intel's Terms and Conditions of Sale for such products, Intel assumes no liability whatsoever, and Intel disclaims any express or implied warranty, relating to sale and/or use of Intel products including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright or other intellectual property right. Intel products are not intended for use in medical, life saving, or life sustaining applications. Intel may make changes to specifications and product descriptions at any time, without notice.

The Intel C440GX+ server board may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

TABLE OF CONTENTS

1. ACPI Power Management 1
  1.1 ACPI states 1
    1.1.1 ACPI Global State Definitions 1
    1.1.2 ACPI Sleeping State Definitions..3
    1.1.3 ACPI Processor Power State Definitions..4
    1.1.4 Device Power State Definitions..5
    1.1.5 Dual processor ACPI States and approximate power consumption 7
  1.2 ACPI - PIIX4/BMC interaction 7
    1.2.1 Power on events..8
    1.2.2 ACPI - PIIX4 Wake/Resume events 8
    1.2.3 G1/S1 resume Events..8
    1.2.4 ACPI - System Control Interrupt (SCI)..9
    1.2.5 Thermal Throttle 9

TABLES and FIGURES

Table 1-1. Summary of Global Power States..3
Table 1-2. Summary of Device Power State..6
Table 1-3. Dual processor Power Consumption during ACPI Global States..7
Table 1-4. C440GX+ server board ACPI State Transitions..7

1. ACPI Power Management

The Advanced Configuration and Power Interface (ACPI) is the key element in implementing Operating System Power Management (OSPM). ACPI is intended for industry wide adoption in order to encourage hardware and software vendors to implement ACPI-compatible (and, thus, OSPM-compatible) systems**.**

The C440GX+ server board will be ACPI compliant as defined by the ACPI 1.0 and the PC98 specifications, ACPI covers more than just power management, but it is only the power management aspect of the specification that affects the hardware implementation. Some of the other goals of ACPI are listed below and have an effect on BIOS, device design and device drivers.

1. Enhance power management functionality and robustness.
   - ⇒ Power management policies too complicated to implement in a ROM BIOS can be implemented and supported in the OS, allowing inexpensive power managed hardware to support very elaborate power management policies.
   - ⇒ Gathering power management information from users, applications, and the hardware together into the OS will enable better power management decisions and execution.
   - ⇒ Unification of power management algorithms in the OS will reduce opportunities for miscoordination and will enhance reliability.
2. Facilitate and accelerate industry-wide implementation of power management.
   - ⇒ OSPM and ACPI will reduce the amount of redundant investment in power management throughout the industry, as this investment and function will be gathered into the OS. This will allow industry participants to focus their efforts and investments on innovation rather than simple parity.
   - ⇒ The OS can evolve independently of the hardware, allowing all ACPI-compatible machines to gain the benefits of OS improvements and innovations.
   - ⇒ The hardware can evolve independently from the OS, decoupling hardware ship cycles from OS ship cycles and allowing new ACPI-compatible hardware to work well with prior ACPI-compatible operating systems.
3. Create a robust interface for configuring motherboard devices.
   - ⇒ Enable new advanced designs not possible with existing interfaces.

1.1 ACPI states

Under ACPI, a system can be defined to be in one of several power states. Power states are used to define the level of power savings and the latency of waking up the system to a fully on condition.

1.1.1 ACPI Global State Definitions

G0 - Working:

A computer state, where the system dispatches user mode (application) threads and they execute. In this state, devices (peripherals) are dynamically having their power state changed. The user will be able to select (through some user interface) various performance/power characteristics of the system to have the software optimize for performance or battery life. The system responds to external events in real time. It is not safe to disassemble the machine in this state.

C440GX+ server board implications:

Full power to the system. The OS may halt one or both of the Processors by issuing a HALT instruction. BIOS should program the Processor to achieve the lowest possible power mode when in at a HALT instruction. The OS can also put embedded devices and plug-in controllers into a lower power mode if device driver allows this. OS idle policy could be set up to spin disk drives down to save additional power after long periods of idleness. This would increase the recovery latency if disk activity were required.

G1 - Sleeping:

A computer state where the computer consumes a small amount of power, user mode threads are *not* being executed, and the system “appears” to be off (from an end user’s perspective, the display is off, fans stop running etc.). To meet PC98 requirements, mandated by Microsoft, the system must support a G1 sleeping state. To do this, the system has to support at least one of the following sleep states: S1, S2, or S3.   The S4 and S4 BIOS sleeping states are NOT sufficient to meet the SDG 1.0 requirement. The "S" sleep states will be discussed in greater detail in the next section.

The latency for returning to the Working state varies on the wakeup environment selected prior to entry of this state (for example, should the system answer phone calls, etc.). Work can be resumed without rebooting the OS because large elements of system context are saved by the hardware, the rest by system software. It is not safe to disassemble the machine in this state

C440GX+ server board implications

The C440GX+ server board will support both the S1 sleep state and the S4 sleep state. The server board will provide the means for the PIIX4 to send STPCLK to both processors, blank the console (already present as part of the security screen blank), and power down the system fans. Because the power supply will still be on and the processors will still be dissipating some power, the power supply and processor fans will be left on. The SCSI controller must supply a means for commanding the disk drives to power down, as they are a major consumer of power and a producer of noise. The hardware will provide a signal to the BMC so that the power LED can be blinked to indicate the system is in a sleeping state.

G2/S5 - Soft Off:

A computer state, where the computer consumes a minimal amount of power. No user mode or system mode code is run. This state requires a large latency in order to return to the working state. The system’s context will not be preserved by the hardware. The system must be restarted to return to the working state. It is not safe to disassemble the machine in this state.

C440GX+ server board implications

The C440GX+ server board will support a S5 Soft-off. The system can be manually, or via Software, be turn off without a save to disk

G3 - Mechanical Off:

A computer state that is entered and left by a mechanical means (e.g. turning off the system’s power through the movement of a large red switch). Various government agencies and countries require this operating mode. It is implied by the entry of this off state through a mechanical means that no electrical current is running through the circuitry and it can be worked on without damaging the hardware or endangering the service personnel. The OS must be restarted to return to the working state. No hardware context is retained. Except for the real time clock, power consumption is zero.

C440GX+ server board implications

If the A/C cord is removed from the wall or power supply, 5V Standby will not be present on the server board. The only power remaining is that provided by the on board battery for the RTC function inside the PIIX4.

Table 1-1. Summary of Global Power States

Global System State	Software Runs	Latency	Power Consumption	OS restart required	Safe to disassemble computer	Exit state electronically
G0 – Working	Yes	0	Large	No	No	Yes
G1 – Sleeping	No	>0, varies with sleep state.	Smaller	No	No	Yes
G2/S5 – Soft Off	No	Long	Very near 0	Yes	No	Yes
G3 – Mechanical Off	No	Long	RTC battery	Yes	Yes	No

Note that the entries for G2/S5 and G3 in the Latency column of the above table are “Long.” This implies that a platform designed to give the user the appearance of “instant-on,” similar to a home appliance device, will use the G0 and G1 states almost exclusively (the G3 state may be used for moving the machine or repairing it).

1.1.2 ACPI Sleeping State Definitions

Sleep states (Sx states) are types of sleeping states within the global sleeping state, G1. The Sx states are briefly defined below.

S1 Sleeping State:

The S1 sleeping state is a low wake-up latency sleeping state. In this state, no system context is lost (Processor or chip set). The system context is maintained by the hardware.

C440GX+ server board implications

Both processors are idle and the OS has placed one of the processors into a HALT state. The OS will then have the option of issuing an S1 or S4 sleeping state command. For entering an S1 sleeping state, the processor issues a command to the PIIX4 in order to cause STPCLK_L to be asserted (and optionally STP_L) to both processors thus placing the system into a G1/S1 state. The OS will have to mask off interrupts, in the PIIX4, so that the only way to wake the processor is via one of the predefined ACPI wake events. The OS should be able to blank the console with the exception of a visible LED, on the front panel, that blinks in order to indicate to the user that the system is sleeping. Prior to issuing the S1 command the OS should be able to spin down all disks, and stop all system fans so that, to the user, the system appears to be off. Stopping the fans should only be allowed if the system is not under thermal stress. Memory will only be refreshed. The criterion for an S1 sleeping state is that to the user, the system appears off.

S2 Sleeping State

The S2 sleeping state is a low wake-up latency sleeping state. This state is similar to the S1 sleeping state except that, the processor and system cache context is lost (the OS is responsible for maintaining the caches and Processor context). Control starts from the processor’s reset vector after the wake-up event.

C440GX+ server board implications

This requires the server board to support isolated power planes for processor and cache so that they could be switched on and off. This state will NOT be supported on the C440GX+ server board.

S3 Sleeping State:

The S3 sleeping state is a low wake-up latency sleeping state where the system context is lost with the exception of system memory. Processor, cache, and chip set context are lost in this state. Hardware maintains memory context and restores some processor and L2 cache configuration context. Control starts from the processor’s reset vector after the wake-up event.

C440GX+ server board implications

This requires the server board to support isolated power planes for processor, cache, and chipset so that they could be switched on and off. This state will NOT be supported on the C440GX+ server board.

S4 - Non-Volatile Sleep:

The S4 Non-Volatile Sleep state (NVS) is a special global system state that allows system context to be saved and restored (relatively slowly) when power is lost to the server board. If the system has been commanded to enter the S4 sleeping state, the OS will write the system context to a non-volatile storage file and leave appropriate context markers. The machine will then enter the S4 sleeping state. When the system leaves the Soft Off or mechanical Off state, transitioning to the working (G0) state and restarting the OS, a restore from a NVS file can occur. This will only happen if a valid NVS data set is found, certain aspects of the configuration of the machine have not changed, and the user has not manually aborted the restore. If all these conditions are met, as part of the OS restarting, it will reload the system context and activate it. The net effect for the user is what looks like a resume from a sleeping (G1) state (albeit slower). The aspects of the machine configuration, that must not change, include but are not limited to disk layout and memory size. However, it may be possible for the user to swap a PC Card or a Drive Bay device.

Note that for the machine to transition directly from the Soft Off or Sleeping states to the S4 sleeping state, the system context must be written to non-volatile storage by the hardware. Entering the working state first, so that the OS or BIOS can save the system context, takes too long from the user’s point of view. The transition from mechanical Off to the S4 sleeping state is likely to be done when the user is not there to observe it.

Because the S4 sleeping state relies only on non-volatile storage, a machine can save its system context for an arbitrary period of time (on the order of many years).

C440GX+ server board implications

For entering the S4 sleeping state, much of the same activity will take place, except, the state of all the controllers and memory will be saved to disk before entering the S4 sleeping state. In the S4 sleeping state the power supply is turned off so only the 5V Standby voltage rail will be available to power the system board. All S4 sleeping state resume events have to be powered from 5V Standby. The OS could also be commanded to perform this operation via the front panel sleep button or from the user interface. The system can be manually woken up via the front panel power button, or, via one of the predefined ACPI wake events.

S5 Soft Off State:

The S5 sleeping state is similar to the S4 sleeping state except the OS does not save any context nor enable any devices to wake the system. The system is in the “soft” off state and requires a complete boot when awakened. Software uses a different state value to distinguish between the S5 sleeping state and the S4 sleeping state in order to allow for initial boot operations within the BIOS to distinguish whether or not the boot is going to wake from a saved memory image.

C440GX+ server board implications

A system commanded off with no wakeup events enabled, has to be manually turned back on and the OS rebooted. Memory has not been saved to disk.

1.1.3 ACPI Processor Power State Definitions

Processor power states (Cx states) are processor power consumption and thermal management states within the global working state, G0.   The Cx states are briefly defined below.

C0 Processor Power State:

While the processor is in this state, it executes instructions normally.

C1 Processor Power State

This processor power state has the lowest latency, The hardware latency on this state is required to be low enough that the operating software does not consider the latency aspect of the state when deciding whether to use it. Aside from putting the processor in a non-executing power state, this state has no other software-visible effects.

C440GX+ server board implications

Halt command executed.

C2 Processor Power State:

The C2 power state offers improved power savings over the C1 power state. The worst-case hardware latency for this state is declared in the FACP table and the operating software can use this information to determine when the C1 power state should be used instead of the C2 power state. Aside from putting the processor in a non-executing power state, this state has no other software-visible effects.

C440GX+ server board implications

This state cannot be supported as part of the global system state G0 as it requires the PIIX4 to supply separate STPCLK signals and control registers for each processor. This C2 ‘Stop Clock’ state can only be supported as a by-product of entering a G1/S1 sleeping state.

C3 Processor Power State:

The C3 power state offers improved power savings over the C1 and C2 power states. The worst-case hardware latency for this state is declared in the FACP table, and the operating software can use this information to determine when the C2 power state should be used instead of the C3 power state. While in the C3 power state, the processor’s caches maintain their state but ignore any snoops. The operating software is responsible for ensuring that the L1 and L2 caches maintain coherency.

C440GX+ server board implications

This state cannot be supported as part of the global system state G0 as it requires the PIIX4 to supply separate STP_CPU signals and control registers for each processor.

1.1.4 Device Power State Definitions

Device power states are states specific to particular devices; as such, they are generally *not* visible to the user. For example, some devices may be in the Off state even though the system as a whole is in the Working state.

Device power states apply to any device on any bus. They are generally defined in terms of four principal criteria:

- Power consumption - how much power the device uses.
- Device context - how much of the context of the device is retained by the hardware. The OS is responsible for restoring any lost device context (this may be done by resetting the device).
- Device driver - what the device driver must do to restore the device to full on.
- Restore time - how long it takes to restore the device to full on.

The device power states are defined below. These states are defined very generically here.

Many devices do not have all of the four power states defined above. Devices may be capable of several different low power modes, but if there is no user-perceptible difference between the modes only the lowest power mode will be used. The *Device Class Power Management Specifications*, which are separate from this specification, describe which of these power states are defined for a given type (class) of device and define the specific details of each power state for that device class.

D3 - Off:

Power has been fully removed from the device. The device context is lost when this state is entered so the OS will reinitialize the device when power is restored. Since device context and power are lost, devices in this state do not decode their address lines. Devices in this state have the longest restore times. This state is defined for all classes of devices.

C440GX+ server board implications

All embedded devices on the server board support this device power state, and it is required by ACPI as this is the state devices must be prepared to enter as part of the G1/S4 sleeping state.

D2:

Each class of device defines the meaning of the D2 device state; it may not be defined by many classes of devices. In general, the D2 device state is expected to save more power and preserve less device context than a D1 or D0 device state. Buses in the D2 device state may cause the device to loose some context (i.e., by reducing power on the bus, thus forcing the device to turn off some of its functions).

C440GX+ server board implications

Devices may lose register context so they would have to save and restore information into NVRAM. No special hardware support would need to be provided. None of the embedded controllers on the server board support this device state.

D1:

Each class of device defines the meaning of the D1 device state; it may not be defined by many classes of devices. In general, the D1 device state is expected to save less power and preserve more device context than the D2 device state.

C440GX+ server board implications

No special hardware support is provided. None of the embedded controllers on the server board support this device state.

D0 - Fully-On:

This state is assumed to be the highest level of power consumption. The device is completely active and responsive, and is expected to continuously remember its entire relevant context.

C440GX+ server board implications

All embedded PCI devices on the server board support this device state

Table 1-2. Summary of Device Power State

Device State	Power Consumption	Device Context Retained	Driver Restoration
D0 - Fully-On	As needed for operation.	All	None
D1	D0>D1>D2>D3	>D2	<D2
D2	D0>D1>D2>D3	<D1	>D1
D3 – Off	0	None	Full init and load

Note: Devices often have different power modes within a given state. Devices can use these modes as long as they can automatically switch between these modes transparently from the software, without violating the rules for the current D*x* state the device is in. Low power modes that affect performance, (i.e., low speed modes) or, that are not transparent to software, cannot be done automatically in hardware; the device driver must issue commands to use these modes

1.1.5 Dual processor ACPI States and approximate power consumption

Table 1-3. Dual processor Power Consumption during ACPI Global States

	Processor-state		Memory		Saved to Disk	Disk Spin		Server board Power	System Power	ACPI Mode	Power
G0/S0	Run-C0	56W	Active	26W	N/A	Spin	65W	16W	ON		163
G0/S1	Halt-C1	5W	Refresh	20W	N/A	Spin	65W	12W	ON		102
G0/S1	Halt-C1	5W	Refresh	20W	N/A	Spin down	27W	12W	ON		64
G1/S1	Stopclk-C2	5W	Refresh	20W	NO	Spin	65W	6W	ON		96
G1/S1	Stopclk-C2	5W	Refresh	20W	NO	Spin down	27W	6W	ON		58
G1/S4	OFF	0W	OFF	0W	YES	OFF	0W	0W	5V Standby		0
G2/S5	OFF	0W	OFF	0W	NO	OFF	0W	0W	5V Standby		0
G3/S4	OFF	0W	OFF	0W	YES	OFF	0W	0W	NONE		0
G3/S5	OFF	0W	OFF	0W	NO	OFF	0W	0W	NONE		0

Table 1-4. C440GX+ server board ACPI State Transitions

G0/S0 >	G3/S5 Loss of A/C	> G0/S0 Restore A/C		
G0/S0 >	G2/S5 Power off	> G3/S5, Loss A/C	> G2/S5 Power off	>G0 restore system
G0/S0 >	G1/S4 NV-Sleep, save to disk	> G3/S4 Lose A/C > G0 Wake-On-LAN > G0 Wake on Ring > G0 RTC Wakeup > G0 P/S ON	> G1/S4 restore A/C	> G0 Wake-On-LAN > G0 Wake on Ring > G0 RTC Wakeup > G0 P/S ON
G0/S0 >	G1/S1 Sleeping	> G1/S4 > G0/S0 Interrupt	> G0 Wake-On-LAN > G0 Wake on Ring > G0 RTC Wakeup > G0 P/S ON > G0/S0 Interrupt	
G0/S0 >	G0/C1 (idle Processor)	> G0/C0 Interrupt > G0/C2 Both Idle > G0/C2/Disk Spindown (Optional)	> G1/S1 Both Very Idle, Spindown (Optional) > G0/C0 Interrupt	> G0/C0 Interrupt

1.2 ACPI - PIIX4/BMC interaction

The BMC interacts with the PIIX4 to control the powering on and the powering off of the system. Essentially, all commands to power up and power down the system pass through the BMC and are then sent to the PIIX4. The BMC uses the PIIX4 power button input to accomplish this. Depending on how the PIIX4 has been programmed, the request to power down the system is handled by an ACPI OS, or, by the SMI handler for Non ACPI OS's.   The net effect of a power down request is that the PIIX4 will assert the output SUSC_L. The BMC will look for this signal and shall control the power supply signal PS_ON appropriately. To power up the system, the BMC will assert the power button input to the PIIX4 and wait for the PIIX4 to de-assert the SUSC_L signal. When the signal is de-asserted the BMC will turn on the power supply by using the PS_ON signal. Note that the system can be turned on and off by events that are not under the control of the BMC, therefore, the BMC must only use the SUSC_L signal as the source of a power on/off request.   Note that the BMC when in secure mode should block the generation of the power button signal to the PIIX4 but not block the SUSC_L signal.

1.2.1 Power on events

The following events should be able to cause the PIIX4 to issue a system wake up, even if the OS is a non-ACPI OS.

- Wake-On-LAN header (used by a plug in card).
- A COM 2 Ring indicator can be provided as a build option to replace Wake-On-LAN.
- PCI_PME (used by the embedded LAN controller).
- RTC - Note that if the PIIX4 observes a loss and restore of A/C power, any RTC wakeup command programmed into the PIIX4 will be lost.

The BMC can cause a system Wake up via:

- The SMM port.
- A command on the IMB bus.
- A command over the Emergencyl Management Port (or Ring Indicator if the EMP is disabled).
- A front panel power button.
- A restoration of A/C power if the system was ON prior to losing A/C power.

1.2.2 ACPI - PIIX4 Wake/Resume events

Under ACPI only a specific number of events can cause a legal wake event. Advanced Power Management has a much wider range of wake events but is not supported by C440GX+.

In order to comply with the ACPI wake/resume programming model, the PIIX4 contains two registers that control system wake up. The two registers are the Power Management Resume Enable Register (Base+02h) and the General Purpose Enable Register (base+0Eh).   These two registers map to the ACPI defined registers PM1a_EVT_BLK and GPE0_EN.

Note that any resume event must be able to cause a SCI.

The legal PIIX4 inputs for resume/wake events are RI, LID, GPI1, THRM and PWRBTN#. USB activity can only be used for a S1 resume, not a S4 wakeup. USB activity can only be detected if the system is already powered up as the system provides power to the USB peripherals and the USB logic is not on the PIIX4 resume Power well i.e. on 5V Standby. The RTC can only cause a S4 wake up and not a resume event as it does not cause a SCI. In the G1/S1 state the BMC will not assert PWRBTN as it is already asserted. Therefore, use of RI, RTC and USB for resume events will not work. The BMC will be connected to the LID input so that it can cause a S4 wake up and a S1 resume event.

1.2.3 G1/S1 resume Events

The number of events that can be used to resume to system from a G1/S1 to a G0/S0 state is very limited. Theoretically, all the S4 Wake-up events except USB activity can also be S1 resume events. In practice Wake-On-LAN has been defined as a S4 wake event and thus, normal LAN traffic will not cause a resume event unless the NIC controller can be programmed to cause a PCI_PME_L assertion. This also applies to the modem RI unless the modem board is a PCI modem and supports the assertion of PCI_PME_L on detection of a RI, then it cannot wake up the system.

There is a stretch goal of having the BMC detect RI from COM2 and causing a S1 resume event. The server board is wired to allow this to happen.

1.2.4 ACPI - System Control Interrupt (SCI)

The SCI is used to replace SMI in all possible situations. The Microsoft* Windows NT* Operating System is mandating that it never looses control for more than an ever declining time period. The SCI will have special abilities and can be generated by predefined activity detected within the PIIX4. The SCI comes out of the PIIX4 GPO29 pin (B3) and will be connected to the IRQ9 input of the IOAPIC and PIIX4. The IOAPIC will have to be programmed to look for an active low input. As the signal is active high, IRQ9 cannot be shared with any other ISA device and has to be dedicated to the SCI. In a uni-processor non-APIC based system, the SCI is connected to the internal IRQ9 of the PIIX4. IRQ9 must be programmed for high level sensitive. This is done with the ELCR2 register.

When IRQ9 is being used for the SCI, the signal has to be removed from the ISA slots. This will be accomplished by logic on the server board controlled by a PIIX4 GPIO bit.

NIC Wake-On-LAN will be connected to RI so that it can wake the system up from a S4 state.

The COM2 Ring indicator goes to the BMC so it will be able to wake from a S4 state on a Modem ring via the PWRBTN or wake from a S1 state via the BMC/LID connection.

1.2.5 Thermal Throttle

The BMC will be hooked up to the PIIX4 THERM_L input and can cause a thermal event to be seen by the OS. This is referred to as a runtime event. Once a PIIX4 pin is defined for a run time event, it cannot be used for a wake/resume event. Because of multi-processor issues with Microsoft Windows NT, only a uni-processor ACPI system can implement thermal throttling. Although the system is wired to allow this, an ACPI compliant method has to be implemented by BIOS and Firmware to allow the OS to request the temperature of the processor. This is not yet a commitment as no ACPI OS is available for testing. Other more robust operating systems, that are insensitive to runtime changes in processor speed, can implement thermal throttling either in the OS or in BIOS by enabling the H/W throttling mechanism in the PCNTRL register (base+10).

Thermal throttling is managed by cycling the STPCLK signal, from the PIIX4, on and off. The period of the cycle is set by the PIIX4 and is 244us. The on/off duty cycle can be in increments of 12.5% or approximately 30us. If enabled, hardware thermal throttling will occur, with a fixed duty cycle, if the THRM input is active for 2 seconds. OS throttling can implement variable duty cycles, but it would require assistance from the BMC in obtaining the system temperature.

For Microsoft Windows NT, the BMC will assert THERM whenever any sensor in the system reaches a thermal limit. Microsoft Windows NT will then cause the system to perform a shut down of the system to prevent data corruption. The system administrator should reduce the operating temperature to a safe level before attempting to restart the system.