WHITE PAPER
Intel® Power Management & I350 Ethernet Controller
Network Connectivity

Intel® I350 Ethernet Controller and DMA Coalescing

Intel® Ethernet Power Management Technology with DMA Coalescing enables users to determine how to meet their energy efficiency and operational goals.

"Intel Power Management Technology with DMA Coalescing enables end users to make a range of choices to determine which trade offs are acceptable to meet their operational goals."

Zane Stabley
Int el Corporation

Carl Hansen
Intel Corporation

Introduction

Power consumption is a significant concern for today's data centers. Power is a monthly fixed cost that all data center providers must pass on to their customers. Competitive industry-wide pricing pressure requires that data center providers find intelligent and creative ways to keep power costs down. In addition, regulatory and OpX factors aimed at reducing total energy consumption have created a demand for more energy-efficient computer platforms. Yet, end-users still need the ability to use the peak performance of their assets to meet business objectives. Energy efficiency is not strictly measured by raw peak or idle power consumption. High performance devices operating at maximum performance for short durations, and then returning to a low-power idle state, are typically the most energy efficient configurations. Intel® Ethernet Power Management Technology with DMA Coalescing enables end-users to make a range of choices to determine which tradeoffs are acceptable to meet their operational goals.

Power Management Technology

Intel's Power Management Technology (PMT) is a standards-based solution, leveraging existing ACPI* and PCI* standards, as well as existing platform power management capabilities of the CPU, chipset and operating system.

PMT provides solutions to common power management approaches by:

- Reducing idle power
- Reducing capacity and power as a function of demand
- Whenever possible, operating at maximum energy efficiency
- Enabling functionality only when needed

Reducing Idle Power

With the Intel® I350-based network controllers and adapters, integrated quad-port configurations consolidate and coordinate functionality between ports on the adapter, effectively increasing energy efficiency. The Intel® I350 also supports PCI power management states, which helps to reduce overall power consumption by reducing power when a device is in an idle state.

The I350 also incorporates a high-efficiency integrated switching-voltage regulator (SVR) that reduces overall BOM cost and design complexity. Its design also enables a more efficient power supply to the component.

Reducing Capacity and Power as a Function of Demand

Intel's PMT incorporates IEEE* 802.3az support, (http://www.ieee802.org/3/az/index.html) also known as Energy Efficient Ethernet or EEE.

Studies indicate that the majority of platforms—both client and server—only use a fraction of the available bandwidth of the local link. Ethernet traffic typically

Table of Contents

occurs in bursts, leaving long periods of inactivity. IEEE 802.3az enables the network interface to enter into a Low-Power-Idle (LPI) mode when the adapter detects that the network link is not being fully used. This enables link partners to save energy by cycling between active and LPI states.

Operation at Maximum Efficiency

Intel's PMT provides a new mode of operation called "DMA Coalescing." It changes the system behavior of the LAN interface by changing how frequently packet data is delivered to the system by batching the delivery of packet data and device interrupts to the chipset, CPU and memory.

This behavior has the following effects:

- By batching and increasing the amount of data transferred during any given time to the system, the LAN device enables the rest of the system to enter into low-power platform states, that is PCIe enters ASPM L1, the CPUs activate Package Cx states, and main-memory goes into self-refresh. DMA coalescing enables these components to stay in these low-power platform states for longer periods.
- Intel's PMT attempts to make the DMA frequency predictable. This predictability enables the host CPU to pick a deeper low-power state than it might otherwise choose.
- When the CPU wakes to process network activity, the operating system is able to run at higher efficiency because software has more "work" to do for any given interrupt. The observable effect, with benchmarks, is, with increasing network I/O block sizes, CPU usage drops and I/O bandwidth increases.

Figure 1 shows that without DMA Coalescing the platform is typically kept in higher power states. The vertical lines show the random nature of platform interrupts. Power consumption, represented by the top line, is higher overall because the processor, memory and other system components are brought out of lower power states to handle the incoming data.

In addition, system components are not allowed enough time to achieve deeper low-power states.

Figure 2 shows that, with DMA Coalescing, the incoming data packets and interrupts associated with these DMA calls are intelligently batched to keep the system devices in lower power states. This enables the system to handle the packets and interrupts more efficiently. The technique also gives system components the opportunity to achieve deeper low power states.

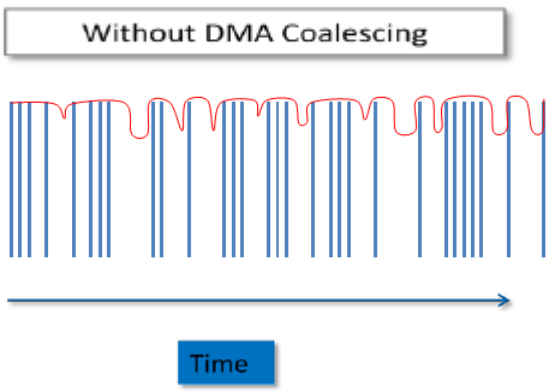

Figure 1

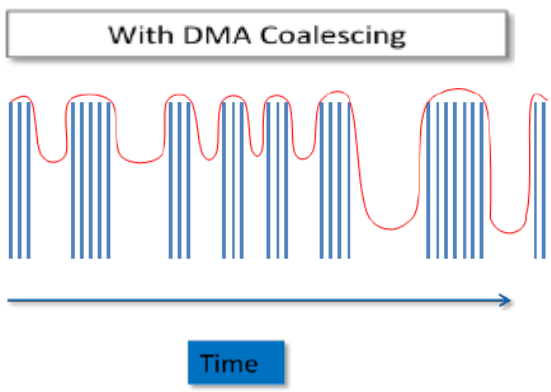

Figure 2

Note: One impact of delaying interrupts and DMA operations is an increase in latency. Most (not all) applications are quite tolerant of latency.

DMA coalescing is accomplished by using the existing transmit and receive buffers on the LAN device to store packets, rather than immediately transferring packet data to or from host memory (as current LAN solutions do). After either a given amount of network data has been buffered (called a watermark), or, after a configurable timer expires, the LAN device exits out of coalescing mode and bursts data accesses and interrupts to the platform. DMA coalescing also enhances previously existing interrupt moderation behavior by throttling the observed device interrupt rate in conjunction with the configurable DMA coalescing timer rate. The interrupt rate is governed by the Interrupt-Moderation-Rate (ITR).

Enable Functionality Only When Needed

With Intel's PMT's support of the ECMA-393 ProxZzzy specification, servers can move to low-power standby states (such as S3), maintain network presence, and be remotely activated via a variety of wakeup packet types.

Intel also supports Low-Power-Link-Up (LPLU). This facility reduces the link power usage in S3 by negotiating the lowest link-speed (where bandwidth capacity isn't required).

DMA Coalescing Experiments & Testing

Experiments were performed to evaluate the power saving benefits of Intel PMTs and the impact on network performance. Intel's PMT scales to reduce power consumption over a wide range of network usage levels. (See Figure 3.)

At network usage below 5%, EEE (802.3az) was most effective, since there is more time to keep the link in a low-powered state. DMA coalescing showed no significant benefit at such low usage rates, since not much data is transferred at those rates.

DMA Coalescing is most effective in the 5% to 35% range, with maximum benefit at 25% usage. Above 35%, power saving benefits decrease. Industry studies report that most servers experience usage rates of 20 – 35%, with only 10-15% of a 1 Gbps link's bandwidth used.

At higher usage, interrupt moderation directly reduces platform power by reducing overall CPU usage. This, combined with the Intel I350's low active power, provides the active system power benefit.

Experiments

- Experiments using an Intel® Urbanna DP platform were run as follows:
  1. Vary the network load
  2. Vary Interrupt Moderation Rate
  3. Measure the platform power
  4. Enable DMA Coalescing and vary the DMA coalescing watchdog time
  5. Fix the Interrupt Moderation Rate (ITR) value
  6. Measure the platform power
- Platform – Test setup
  - 2 x 2.93 GHz Quad-core Xeon® CPUs (X5570)
  - 12 GB (2048 x 6) DDR3 1333MHz memory
  - BIOS defaults – enhanced C-states, C6/ Turbo/HT–enabled
  - I350 development - test adapter
  - Linux* 2.6.32 with the following features enabled; tickless, high_res_ timers, hpet_timer, ondemand CPU governor, Powertop- timer_stats and PCI-ASPM.
  - Manually force ASPM L1 on the network adaptor port.
  - Network connection at 1 Gbps.

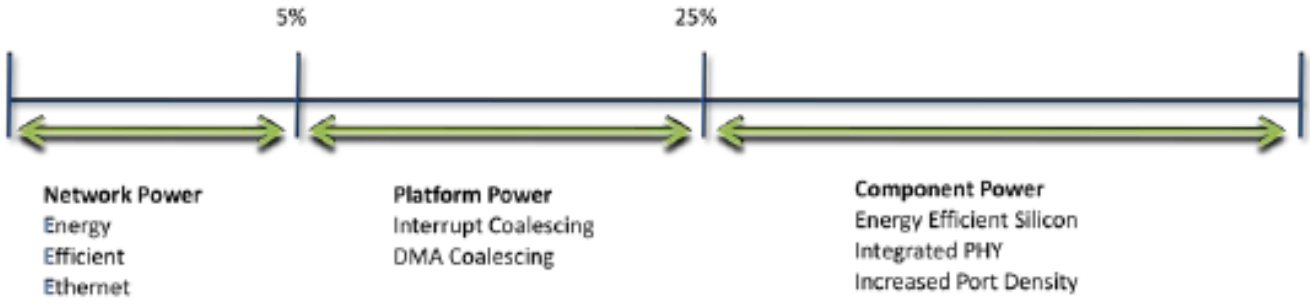

Figure 3

- Set one port as Receive with smartbits = 1514 byte continuous UDP packet stream from another port.

- Results & Observations
  - Throttling interrupts by itself improves power efficiency.
  - Adding DMA coalescing creates further power savings. Figure 4 shows how moderating interrupts improves power efficiency and the addition of DMA coalescing further increases power savings.
  - Peak benefit reached at expected throughput of ~250 Mbps (25%),
  - Beyond optimal throughput, power savings begins to decrease. Figure 4 shows the power savings of a single port using interrupt moderation and DMA coalescing within the context of network usage.
  - DMA moderation benefits increase as more time is allowed for coalescing, for example 250 uS to 5 mS. However, as additional time for coalescing is enabled, response-time latency (if the network data is not sufficient to exceed the device water mark) increases proportionally.
  - Asynchronous activity between two discrete controllers (2x dual-port vs 1x quad-port) will interfere with CPU lower power state entry and duration, reducing DMA coalescing power effectiveness.

Intel® Ethernet I350 Controller

- Integrated Quad Port Silicon
- Intel has achieved DMA Coalescing in an integrated quad port part today!
- Intel synchronizes DMA activity across all four ports of our quad port controllers beginning with the I350

DMA Coalescing Across Multiple Intel Quad Port Adapters

- Through software emulation, Intel is able to synchronize DMA Coalescing between two Intel adapters
- Typical platform power savings of 15 W to 20 W per server with DMA Coalescing enabled on a single four port LAN device
- Additional testing results and details will be forthcoming in future revisions of this document

Additional Configuration Information

The following platform-level configurations and settings dramatically improve the power efficiency of a system using Intel PMT.

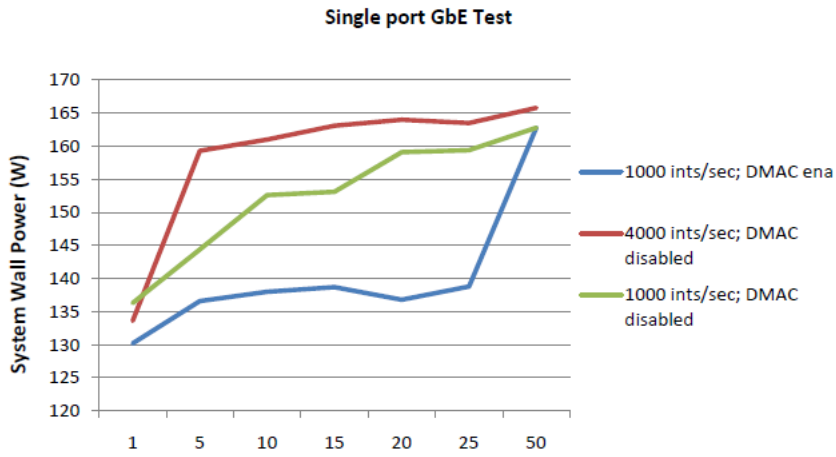

Figure 4

Platform Considerations

Overall, minimize the use of USB* devices. The USB bus is a polled bus; transactions are initiated by the host and not the USB device. Because of this, USB devices contribute more interrupts to the system and make it difficult to control Power Managment. USB 2.0 does support a "suspend" low-power state; but the state's entry/exit latency make it difficult to use effectively. Best results occur when network applications scale across multiple CPU cores as evenly as possible. Enable Receive Side Scaling (RSS) to affinitize interrupts to the CPU cores.

BIOS Tuning

The following settings are typically configured in the BIOS setup screens:

1. Enable C1E, disable C3-report, and enable C6-report

When the OS selects entry into ACPI C3, the BIOS will map this request to the internal CPU C6 state

2. Enable Package C3 and Package C6

This enables the CPU to select, synchronize, and activate a low power mode over multiple CPU cores simultaneously.

3. Enable Enhanced Intel Speedstep Technology (EIST).

Enhanced Intel SpeedStep Technology enables the system to dynamically adjust processor voltage and core frequency. This can result in decreased average power consumption and decreased average heat production.

4. Enable ASPM L1 if possible for additional PCIe power savings.

Software Operating System Tuning

When using Windows* Server 2008 R2:

1. Disable core parking if needed.
2. Install all chipset-specific and device-specific device drivers (such as the Intel® Chipset INF updater, as well as vendor-specific graphics drivers).

Contact your local Intel Field representative to obtain the "SelfTest" tool from http://www.intel.com/cd/edesign/library/asmo-na/eng/434688.htm . The tool verifies the platform BIOS/OS

configuration.

Linux* versions 2.6.33 and later support the required power management hooks to optimize DMA coalescing. Customizations of the kernel enhance the effect:

1. Enable 'tickless' feature with Tick=1000 and preemption mode=Server, CPU idle-Power Management support=enabled.
2. Load CPUFREQ module: cpufreq_ ondemand.
3. If possible, disable PCSCD (Smart Card Daemon).
4. After configuration and boot, run "turbostat" (of powertop version 2.0 or later) to verify 80% or greater Package C3 or Package C6 residency.

Controlling DMA Coalescing Performance

Disabling Interrupt Moderation will also disable DMA Coalescing. DMA Coalescing is disabled by default, but is enabled through the Performance Options tab in the Windows* DMIX interface (Figure 5) and through the command line in Linux. For example:

```
modprobe igb [<option>=<VAL1>,<VAL2>,...]
```

See Table 1. The default value for each parameter is generally the recommended setting, unless otherwise noted.

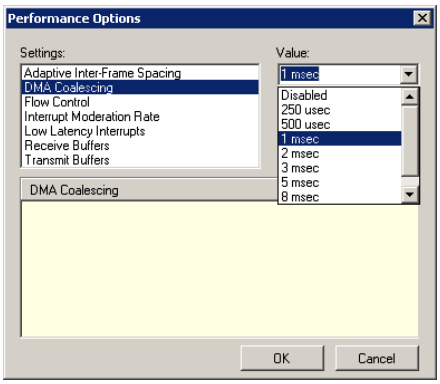

Figure 5

DMAC	0, 250, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000	0 (disabled)	Enables or disables DMA Coalescing feature. Values are in usec's and increase the internal DMA Coalescing feature's internal timer. DMA (Direct Memory Access) allows the network device to move packet data directly to the system's memory, reducing CPU usage. However, the frequency and random intervals at which packets arrive do not enable the system to enter a lower power state. DMA Coalescing enables the adapter to collect packets before it initiates a DMA event. This may increase network latency but also increases the chances that the system will enter a lower power state. Turning on DMA Coalescing may save energy with kernel 2.6.32 and later. This will impart the greatest chance for your system to consume less power. DMA Coalescing is effective in helping potentially saving the platform power only when it is enabled across all active ports. InterruptThrottleRate (ITR) should be set to dynamic. When ITR=0, DMA Coalescing is automatically disabled.

Table 1

The DMA Coalescing max-wait time is adjustable through the same interfaces.

Methods to Verify Behavior

There are three methods to verify behavior: at the PCIe Bus Analyzer level, via Package Cx state residency counters, and raw wall-power measurements of the system.

PCIe Bus Analyzer Method

The PCIe bus analyzer method requires instrumentation of the LAN adapter with a PCIe interposer and capturing PCIe traffic while the device is being used. (A full description of what is required to do this is beyond the scope of this document.) After the traces have been captured, using the analyzer-specific software, visualize the bus usage based on the PCIe transactions. An example graph generated by LeCroy PETracer* appears in Figure 6.

Wall-power Method

The raw wall-power measurement of the system is relatively straightforward with the correct wall-power measurement equipment—such as a Watts Up*, Kill a Watt*, or an equivalent power measurement tool. In general, follow the platform measurement guidelines published in the EnergyStar* standards—specifically allowing for a settling time after the system boots to allow startup processes to complete and then go idle.

Package Cx State Residency Method

The Package Cx state residency method requires special software, as well as a basic background on "C states" on CPUs. "C states" in ACPI corresponds to various CPU functional states, similar to what D-states are for I/O devices, and S-states are for platforms. For example, C0 means the CPU is fully operational and executing instructions. The various higher C-states, such as C1, C2, and C3, correspond to lower and lower power states with longer and longer resume times.

The OS typically requests entry into one of these C states based on its own internal heuristics, as well as an exit latency table provided by the BIOS to the OS.

The BIOS maps processor-specific C-states to the OS-exposed C-states. For example, if the OS calls the ACPI "C1" state, this would likely be mapped to the Intel C1E power state (where the CPU, on exit, resumes execution at the lowest-frequency available, if Enhanced Intel Speed Step, EIST, is also enabled). If the OS invokes "C3," the BIOS could activate either C3 or C6 depending upon how the BIOS is configured. Lastly, although the BIOS may request the CPU go into "C6," the CPU may auto-demote, or select a different, more shallow C state such as C3, based on device access and interrupt delivery patterns.

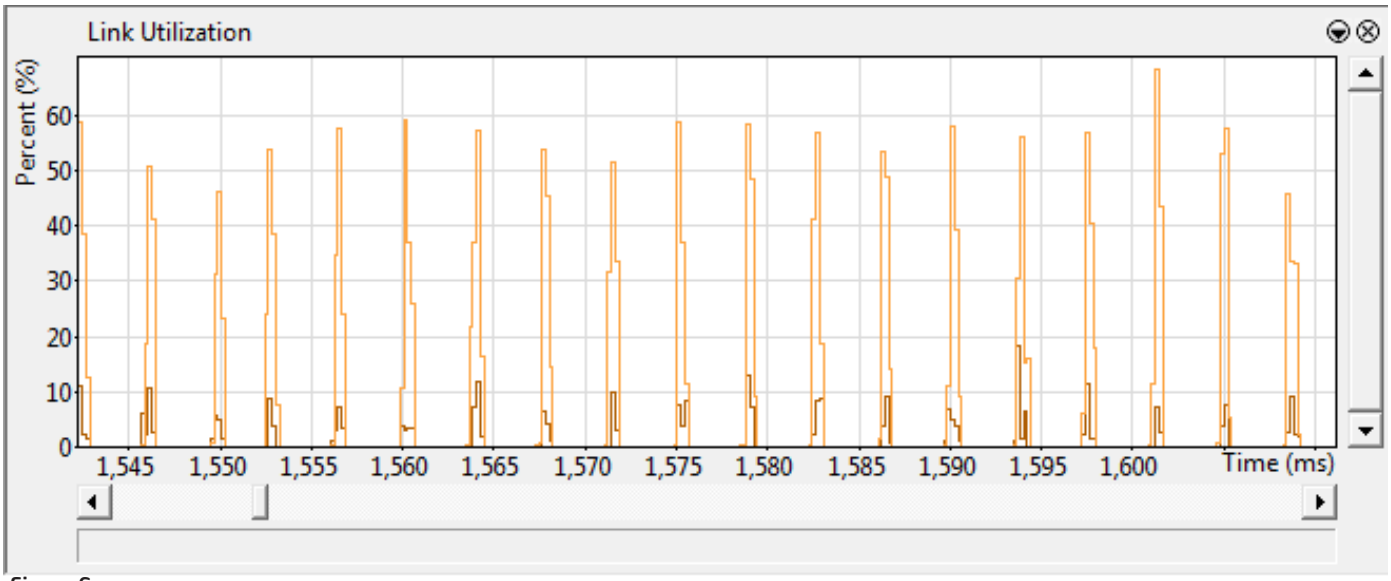

Figure 6

C-state entry by the OS is on a per-core basis. These are called "Core C States" (or CC3, for example, for Core C3). When all CPU cores simultaneously go into the same or deeper C state, the supervisory firmware controlling the package (previously called "socket") can enable the entire package to transition to a Package C state. The dramatic power savings occur whenever the entire package enters these Package C3 or Package C6 states. However, I/O activity, even at seemingly platform idle, such as graphics DMA, disk DMA, or LAN DMA, prevents the package from entry into these deep power states. As such, much of the time spent on platform tuning is used to identify the source of this activity. An example of this is running a copy of Windows that hasn't been activated; Windows causes a small amount of disk activity that isn't noticeable by looking at CPU usage alone.

To determine the current Package C-state residency, special software must be used. On Linux, the powertop* tool versions 2.0 and later support reporting these metrics, as do the Linux "turbostat" tool. For Windows, there are Intel tools: the Intel Battery Life Analyzer tool (BLA) reports the Package C states (as well as the actual cause of the I/O activity in many cases) on Intel client platforms. On server platforms, a special perfmon DLL must be installed to read the processor-specific performance counters.

Ideally, on a well-tuned platform at idle, the platform should see 85% or greater Package C6 % residency. At this point, various networking benchmarks can be stated to evaluate the benefits of DMA coalescing and other Intel PMT features.

References.

Designing Power-Friendly Devices (Intel Whitepaper)

Energy-Efficient Platforms/Green Hill Software (Intel Whitepaper)

Intel® I350 Quad-/Dual-Port GbE LAN Controller Datasheet

For more information on Intel® PMTs and the Intel® I350A, visit **www.intel.com/go/ethernet**

Printed in USA ZS/CH/SU Please Recycle 324826-002US