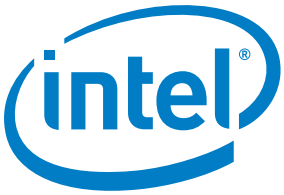


Data Plane Development Kit Power Optimization on Advantech* Network Appliance Platform

White Paper

December 2015

Document Number: 333552-001US

Legal Disclaimer

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described 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.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting: http://www.intel.com/design/literature.htm

Intel, Intel Core, Celeron, Pentium, Xeon , PCM, and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

Contents

Figures

Tables

1.0 Executive Summary

Power management techniques provided by a hardware equipment manufacturer for reducing energy consumption can provide significant opportunities for operational cost savings and other business value. To maximize effectiveness, a combination of hardware BIOS and software optimization needs to be analyzed to allow decision makers to make the right decision in the approach toward “energy efficient”.

This white paper looks into Intel® Core™ i7 processor on Advantech* FWA-4231 network appliance platform’s performance and energy efficiency implications of BIOS tuning options available on the platform for network application workloads. The scope includes analysis of the system level configuration impact on platform power consumption and also delves deeper into software level optimization provided by DPDK, which significantly reduces CPU power consumption thus reduce platform level power.

This white paper evaluates Intel® Architecture (IA)-based platforms and their network application workloads by better understanding the implications of hardware, BIOS, and software design decisions which can lead to great differences in power consumption.

1.1 Terminology

Table 1. Terminology

Term	Description
DPDK	Data Plane Development Kit
DDIO	Intel® Data Direct I/O (Intel® DDIO)
DiffServ	Differentiated Services Computer Networking Architecture
IOU	IO Unit
IPv4	Internet Protocol version 4
L3fwd	L3 Forwarding Sample Application
LLC	Last Level Cache
LPM	Low Power Mode
NGPTIM	Next Generation Polymer TIM
OEM	Original Equipment Manufacturer
PCH	Platform Controller Hub
PCU	Power Control Unit
PCM	Intel® Performance Counter Monitor (Intel® PCM)
PMD	Source Code Analyzer
SKU	Stock Keeping Unit
TIM	Thermal Interface Model

1.2 Reference Documents

Table 2. Reference Documents

Document	Document No./Location
Desktop 4th Generation Intel® Core™ Processor Family, Desktop Intel® Pentium® Processor Family, and Desktop Intel® Celeron® Processor Family - Datasheet – Volume 1 of 2	328897
Desktop 4th Generation Intel® Core™ Processor Family, Desktop Intel® Pentium® Processor Family, and Desktop Intel® Celeron® Processor Family - Datasheet – Volume 2 of 2	328898

1.3 Advantech* FWA-4231 Network Appliance Platform

Figure 1. Advantech* FWA-4231 Network Appliance

The Advantech* FWA-4231 2U Network Appliance can scale from entry-level performance using Intel® Celeron® and Pentium® processors to the mid-range segment using Intel® Xeon® series processors, offering an outstanding price/performance ratio with each Stock Keeping Unit (SKU). In addition to supporting the latest Intel® microarchitecture formerly known as Haswell enhancements for increased CPU performance virtualization support and I/O throughput, the Advantech* FWA-4231 adds support for the latest Gigabit Ethernet controllers, PCIe3* connectivity, remote management, and advanced LAN bypass to create state-of-the-art platforms for specific enterprise networking applications.

The scalability of the Advantech* FWA-4231 positions it ideally for Original Equipment Manufacturers (OEMs) designing high bandwidth systems in telecommunications and enterprise networking. It is ideal for applications in service provider networks for enhanced security, deep packet inspection, and acceleration and subscriber-based services.

Figure 2. Advantech* FWA-4231 Platform Specification

Processor System	CPU	Intel® Xeon® Processor E3-1275/E3-1225/E3-1268LV3 4th Gen Intel® Core™ Processors i5-4570TE, i3-4360, i3-4330, i3-4330TE, Intel® Pentium® Processor G3320TE and Intel® Celeron® Processor G1820TE(LGA1150)	
	LLC	8M/6M/4M/3M/2M	
Chipset	PCH	Intel® C226 PCH	
Memory	DIMM sockets	4	
	Technology	Dual channel DDR3 1333/1600 MHz ECC Un-buffered memory	
	Capacity	Up to 32 GB	
PCIe	Expansion Slots	1 x PCI-Express Gen.3 x8 FH/HL Expansion Slot	
Ethernet	LAN on Board	2 x Intel I210-AT 10/100/1000 Mbps Ethernet	
	NMC modules	4 x NMC modules with PCIex8 gen.3 interfaces Maximum 8GbE ports or 2 x 10GE ports. Please refer to the "Recommended NMC Module List" section for a list of currently available NMCs	
Storage	SATA	4 x 2.5" Swappable 2 x 3.5" internal SATA HDD(optional)	
	Flash	1 x mSATA socket , support Full size/half size module CF or CFast module support (Optional)	
Management Ports & Peripherals	USB	2 x USB2.0 Type A connectors on the front 1 x USB3.0 pin header	
	Serial	1 x RS232 Console port (RJ-45 connector)	
	LCD Module	16 x 2 graphic display, 5 buttons	
	VGA	1 x VGA pin header	
	TPM	Optional	
Power Supply	Wattage	350W redundant AC PSUs (redundant DC PSUs on request) 300W Single AC PSU	
	Input	AC 100 ~ 240 V @ 50 ~ 60 Hz, full range PMBus support on redundant PSU	
Environment		Operating	Non-Operating
	Temperature	0 ~ 40° C (32 ~ 104° F)	-40 ~ 60° C (-40 ~ 140° F)
	Humidity	5 ~ 85 %@ 40° C (104° F)	5 ~ 95 %
Physical	Dimensions (W x H x D)	430 x 88 x 550 mm, 16.6" x 3.4" x 21.6"	
	Weight (N.W)	15kg (33lb)	

1.4 4th Generation Intel® Core™ i7 Desktop Processor

The Desktop 4th Generation Intel® Core™ processor family , Desktop Intel® Pentium® processor family, and Desktop Intel® Celeron® processor family are 64-bit, multi-core processors built on 22-nanometer process technology. The processors are designed for a two-chip platform consisting of a processor and Platform Controller Hub (PCH). The processors are designed to be used with the Intel® 8 Series chipset.

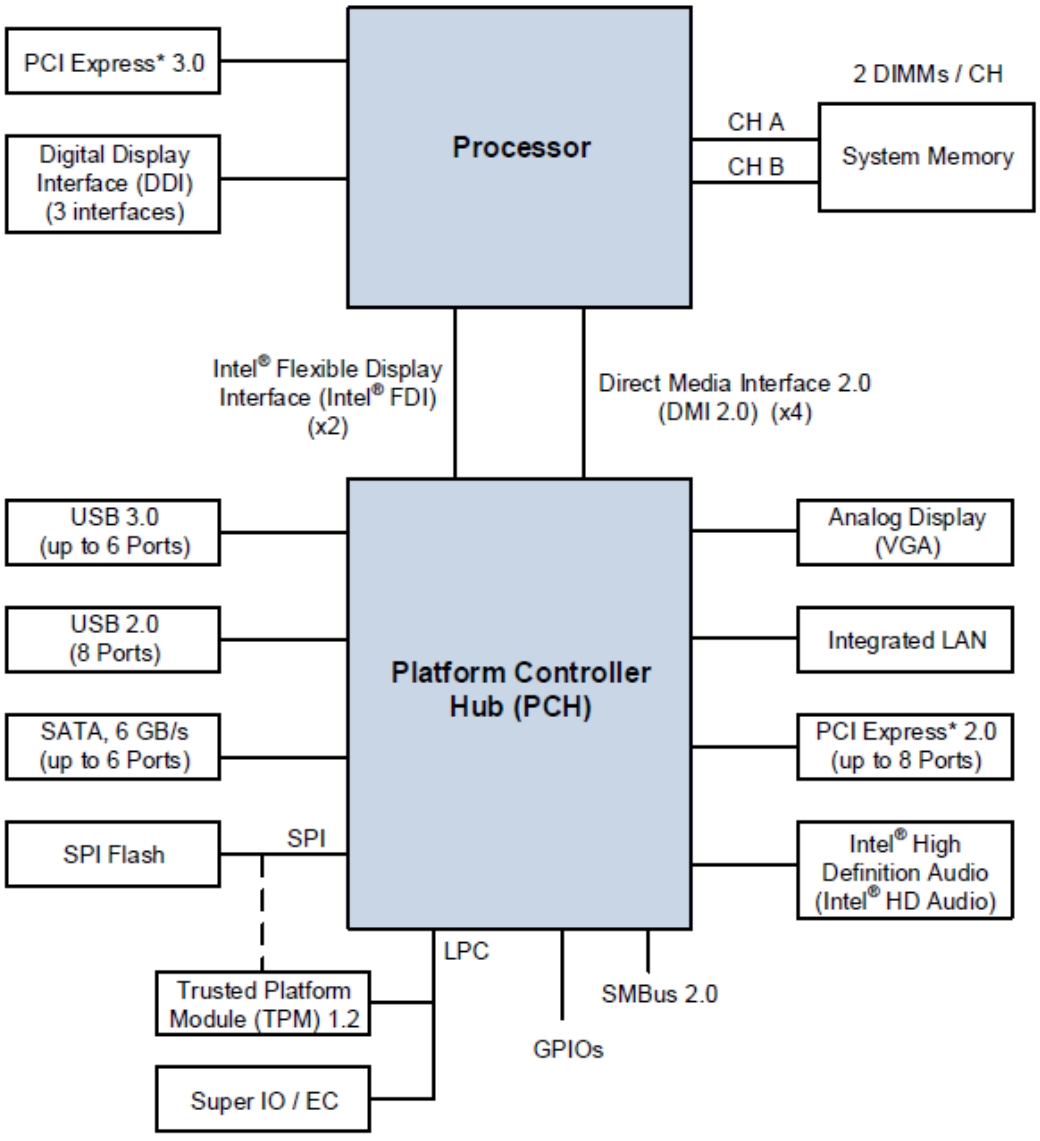
The 4th Generation Intel® Core™ i7 architecture is specifically designed to optimize the power savings and performance benefits from the move to FinFET (non-planar, "3D") transistors which use a 22 nm process.

The 4th Generation Intel® Core™ i7 is the first to have additional capacitors used for a smooth power delivery to the die. The other major improvement is the thermal design, in particular, the Thermal Interface Material (TIM). Intel says a new next-generation polymer TIM (NGPTIM) is now used that should give a greater amount of thermal headroom for overclocking.

At the micro-architecture level, Intel added more power gating and low power modes to the 4th Generation Intel® Core™ i7. The additional power gating gives the power control unit (PCU) more fine grained control over shutting off parts of the core that are not used.

The 4th Generation Intel® Core™ i7 can also transition between power states approximately 25% faster than Intel® microarchitecture code name Ivy Bridge, which lets the PCU be a bit more aggressive in which power state it selects since the penalty of coming out of it is appreciably lower. It's important to put the timing of all of this in perspective. Putting the CPU cores to sleep and removing voltage/power from them even for a matter of milliseconds adds up to the sort of savings necessary to really enable the sort of always-on, always-connected behavior the 4th Generation Intel® Core™ i7 based systems are expected to deliver.

Figure 3. Intel® Core™ i7 -4790 High Level Platform Diagram

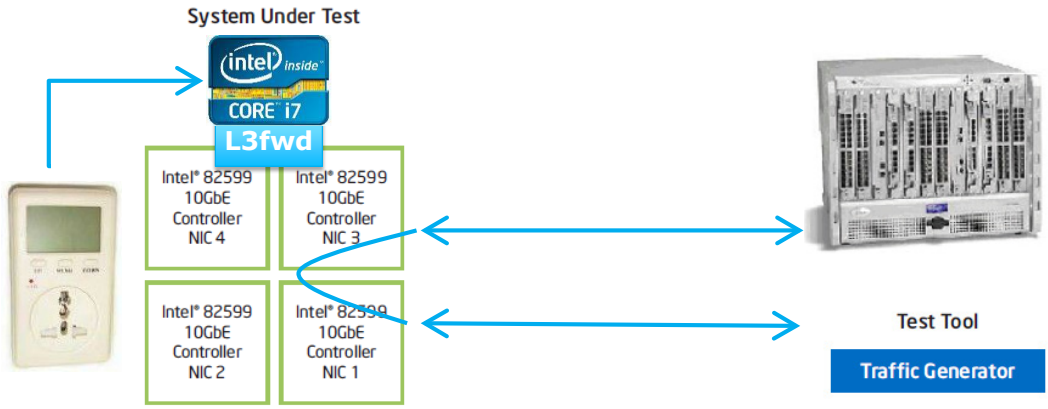


- **GHz per dollar:** For pure Gigahertz speed for the money, the 2011 and 1150 i7s come out on top every time, making them the best value for single threaded applications. For example, a 4th Generation Intel® Core™ i7-4790 running at 3.6 GHz retails for around $300. The comparable quad core Intel® Xeon® running at that clock speed will cost about $50 more.
- **On board graphics:** Intel® Core™ i7 and Intel® Core™ i5 processors all come with onboard graphics, meaning a discrete video card is not required for video display, whereas Intel® Xeon®processor-based PCs cannot be configured without discrete video. Though a discrete card is recommended for anything beyond the most casual gaming or video work, on board graphics are suitable for many home office uses.

1.5 Test Setup

Figure 4 depicts the complete setup of device under test, traffic generator, and power meter used to conduct all our test case. The system is configured with four Intel® 82599 10 Gigabit Ethernet Controllers, on 2 x NMC-100 4E (2 x 10 Gbps, with Intel® 82599).

Figure 4. Test Setup Diagram



1.5.1 Use Case

The primary objective of this white paper is to understand and analyze hardware and software power management features which apply to an actual commercial network appliance platform. The tests are focus around power consumption in an Advantech* commercial network appliance platform, with variation on BIOS configuration, CPU core utilization using software configuration, and network traffic pattern.

The result of the tests are intend to deliver a best/optimize configuration in term of BIOs setting, software, and traffic model.

1.5.2 Test Methodology

The test durations were 30 minutes; the total kw/h from the power meter was collected. The accumulative power consumption can be more accurate than the power drawn by the system on an instance. Both tests were conducted on the same Advantech* FWA-4231 Network Appliance Platform.

1.5.3 Test Application: L3fwd

The application was L3 Forwarding application (l3fwd) and l3fwd-power from Data Plane Development Kit (DPDK) v2.0.

The l3fwd is a simple example of packet processing using the DPDK. The application performs L3 forwarding. The application demonstrates the use of the hash and low power mode (LPM) libraries in the DPDK to implement packet forwarding. The initialization and run-time paths are generic DPDK initialization. The most important port is the forwarding engine which runs 100% of the CPU. First the forwarding decision is made based on information read from the input packet. The lookup method is either hash-based or LPM-based and is selected at compile time.

When the selected lookup method is hash-based, a hash object is used to emulate the flow classification stage. The hash object is used in correlation with a flow table to map each input packet to its flow at runtime. The hash lookup key is represented by a differentiated services (DiffServ) 5-tuple composed of the following fields read from the input packet:

- Source IP Address
- Destination IP Address
- Protocol, Source Port
- Destination Port

The ID of the output interface for the input packet is read from the identified flow table entry.

When the selected lookup method is LPM-based, an LPM object is used to emulate the forwarding stage for Internet Protocol version 4 (IPv4) packets. The LPM object is used as the routing table to identify the next hop for each input packet at runtime. The LPM lookup key is represented by the Destination IP Address field read from the input packet. The ID of the output interface for the input packet is the next hop returned by the LPM lookup.

In general, the DPDK executes an endless packet processing loop on dedicated IA cores that include the following steps:

1. Retrieve input packets through the PMD to poll Rx queue
2. Process each received packet or provide received packets to other processing cores through software queues
3. Send pending output packets to transfer queue through the source code analyzer (PMD)

For hash-based loop up, the forwarding loop is optimized for continuous 4 valid ipv4 and ipv6 packets, they leverage the multiple buffer optimization to boost the performance of forwarding packets with the exact match on the hash table.

As for an LPM lookup, the forwarding loop will match the longest prefix (destination IP) to an output port from the LPM object.

1.5.4 Test Application: L3fwd-power

The L3 Forwarding with Power Management application (L3fwd-power) is an example of power-aware packet processing using the DPDK. The application is based on existing L3 Forwarding sample application, with the power management algorithms to control the P-states and C-states of the Intel processor via a power management library.

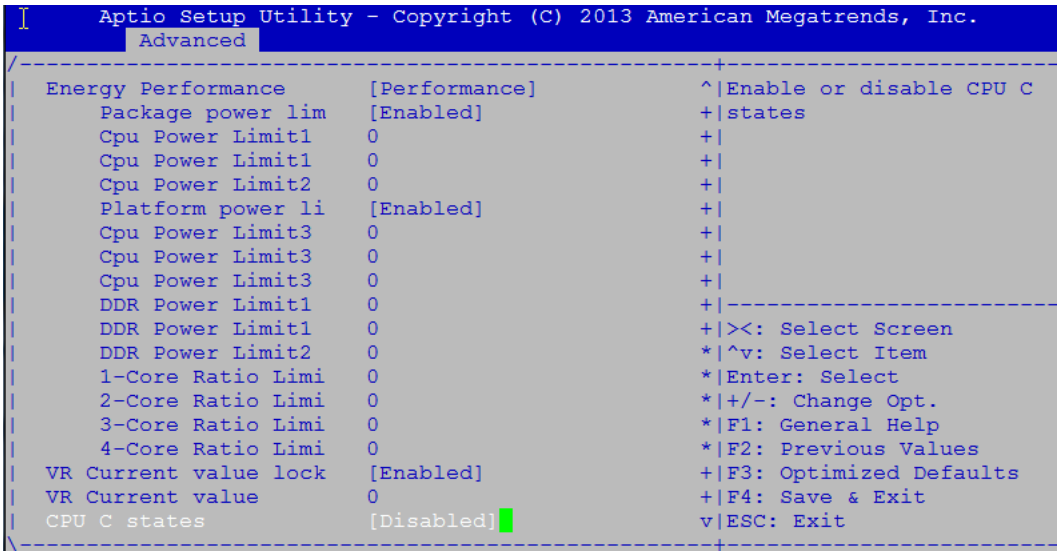
Revisiting DPDK's simple receive then lookup, and transmit the packets out. This process is running constantly even while there are no incoming packets. During the period of processing light network traffic, which happens regularly in communication infrastructure systems due to well-known "tidal effect", the PMD is still busy waiting for network packets, which wastes a lot of power.

This application includes a P-state power management algorithm to generate a frequency hint to be sent to CPUFreq. The algorithm uses the number of received and available Rx packets on recent polls to make a heuristic decision to scale frequency up/down. Specifically, some thresholds are checked to see whether a specific core running an DPDK polling thread needs to increase frequency a step up based on the near to full trend of polled Rx queues. Also, it decreases frequency a step if packet processed per loop is far less than the expected threshold or the thread's sleeping time exceeds a threshold.

1.5.5 BIOS Setting

The platform power profile in the CPU's advanced setting was so that Performance or Balance Performance were used to conduct the test.

Figure 5. BIOS Configuration Setting



1.5.6 Test Case 1 – Comparison between Standard l3fwd vs l3fwd-power

Test Case 1 serves as a baseline test case in which power consumption is collected for the application that will consume the highest power (l3fwd). Then the same was conducted with the same setting on the same hardware.

From this test, the consumption for l3fwd and l3fwd-power should be identifiable; the difference was considered the baseline power savings from l3fwd-power.

```
L3fwd

sudo ./build/l3fwd -c 0xFF -n 4 -- -p 0x03 --
config="(0,0,0),(0,1,1),(0,2,2),(0,3,3),(1,0,4),(1,1,5),(1,2,6),(
1,3,7)"
```

```
L3fwd-power

sudo ./build/l3fwd-power -c 0xFF -n 4 -- -p 0x03 --
config="(0,0,0),(0,1,1),(0,2,2),(0,3,3),(1,0,4),(1,1,5),(1,2,6),(
1,3,7)"
```

Table 3. Power Consumption of L3fwd vs L3fwd-power

Load	0%	10%	30%	50%	70%	90%	100%
Raw RX rate (Mpps)	0	1.49	4.46	7.44	10.41	13.02	14.88
L3fwd (watt)	150.6						153.6
l3fwd-power (watt)	78.9	88.4	101.4	107.3	111.6	113.2	113.6

1.5.7 Test Case 2 – L3fwd-power Power Consumption with Different Packet Sizes

Test Case 2 was conducted to review the effects of different packet sizes on the power consumption of L3fwd-power. The setup for Test Case 2 is similar to Test Case 1 with the only variation being of packet sizes with the data rate running at 100% line rate is 20 Gbps.

Table 4. Power Consumption of L3fwd-power with Different Packet Sizes

Packets Sizes (Bytes)	64	128	256	512	768	1024	1280	1518
Raw RX rate (Mpps)	14.88	8.45	4.53	2.35	1.59	1.20	0.96	0.81
l3fwd-power (watt)	113.6	110.1	103.4	92.6	89.6	88.6	87.4	86.8

1.5.8 Test Case 3 – L3fwd-power Power Consumption with Different BIOS Power Management Settings

Test Case 3 was conducted to review the effects of different power management setting in the BIOS on the power consumption of L3fwd-power. There are a few power profile options available:

- **Maximum Performance**: Provides the highest performance and lowest latency. Use this setting for environments that are not sensitive to power consumption.
- **Balanced Performance (default):** Provides optimum power efficiency and is recommended for most environments.
- **Power Savings Mode**: Provides power savings for environments that are power sensitive and can accept reduced performance.

These profiles are arranged in the order from high to low power consumption. The test setup is similar to the previous test for both the hardware and software environment with the difference of only the power profile used during the measurement with an increase data rate from 10% to 100% line rate of 20 Gbps, and 64 bytes packet size.

Table 5. Power Consumption of L3fwd-power with Different Packet Sizes

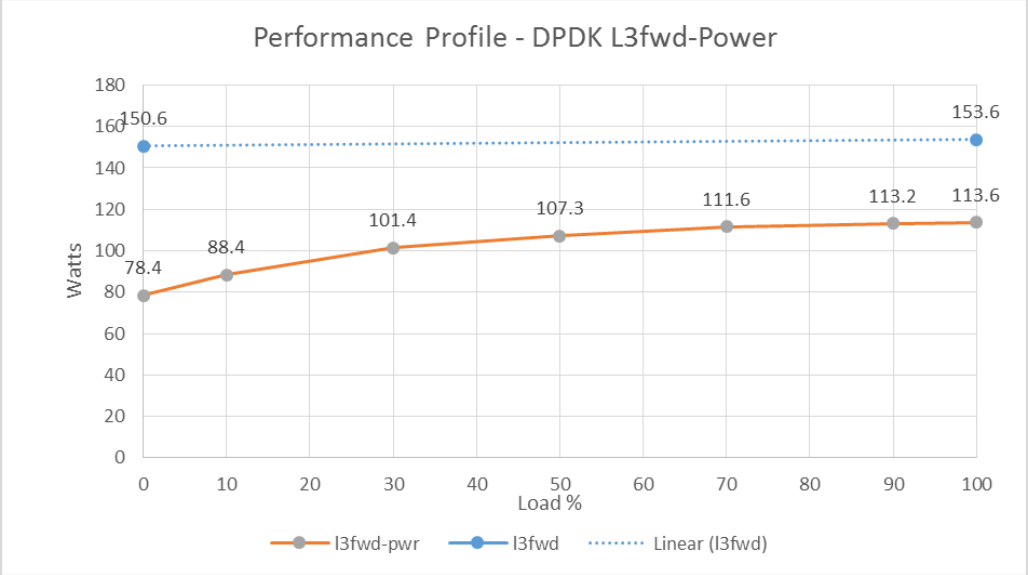
Load	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Raw RX rate (Mpps)	0	1.49	2.98	4.46	5.90	7.44	8.93	10.41	11.57	13.02	14.88
Max Saving	78.9		89.2		95.8		99		99.2		103.8
Perfor-mance	78.9	88.4		101.4		107.3		111.6		113.2	113.6

§

2.0 Analysis

After running a baseline testing of both l3fwd and l3fwd-power, we can now compare their power consumption (Figure 6):

Figure 6. DPDK L3fwd vs L3fwd-power Power Consumption on Different Traffic Loads



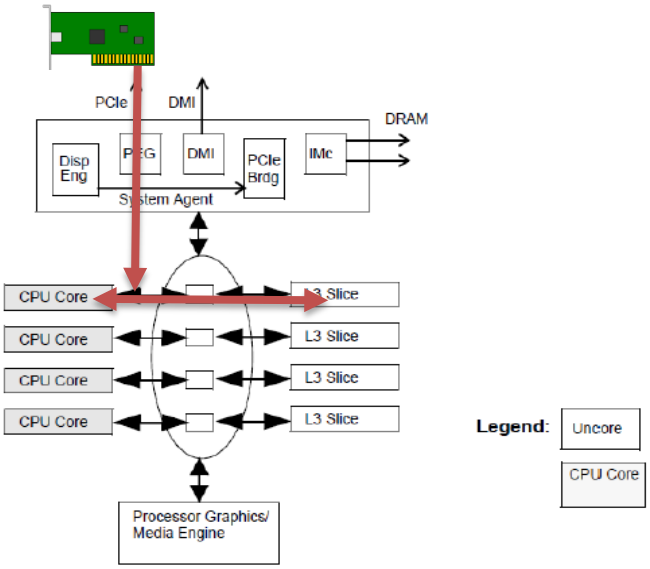
Since Test Case 1 and Test Case 2 setups run using the same parameters and the same amount of cores, memory and uses the same PCIe* network card, the packet transfer rate are 20 Gbps with 64 bytes packets size and we can see that DPDK can receive about 20 Gbps, 64 bytes line rate traffic.

Figure 6 above shows both l3fwd and l3fwd-pwr being executed in identical environments on the same hardware. They are run using the same parameters with the same amount of cores and memory and they use the same PCIe* network card. The packet generator is transmitting at 20 Gbps with 64 bytes packets size. Raw throughput received back by the packet generator is at about 20 Gbps too.

Looking at l3fwd-power consumption, it increased from 150.6 to 153.6 watts from no traffic to 100% 20 Gbps line rate, a total of 3 watts were consume by the traffic. This is due to the traffic entering the system, exercising the PCIe* bus, then use the IO unit (IOU) to enter the core and consume CPU cycles. This is where the 3 watts were used.

Findings are confirmed by using an Intel® Xeon® E5-2600v3 CPU running Intel® Performance Counter Monitor (Intel® PCM). There is a lack of test registers for Intel® PCM to probe for data from either in the memory controller or PCIe* controller. On the Intel® Xeon® E5-2600v3 CPU, there was no increase in memory bandwidth utilization, but the PCIe* transaction counter jumped from 20 k to about 34 M when 64 bytes 20G bps traffic is streamed into the system. The packets are directly written into the last level cache (LLC) using Intel® Data Direct I/O (Intel® DDIO). Figure 7 illustrates the part of CPU components utilized in this test.

Figure 7. Core-i7 Core Components Supporting Network Traffic



Comparing l3fwd and l3fwd-power, their recorded idle power while both application are running are 78.4 watts and 150.6 watts respectively. The difference is about 72.2 watts which is 47.9% below the standard l3fwd. This power saving is enormous. This number is only the power used by system itself. We have NOT includes the energy consumption associated with the extensive space cooling load imposed, and the power loss for delivering the power in a data center/enterprise to the system.

In fact, based on Emerson's Energy Logic model, they have determined that every one watt saved at the server component translates into a savings of over 2.84 watts in power and cooling; .this means that a 205 watt of total power savings is getting closer. Furthermore the system power consumption when loaded at 100% is about 40 watts.

Figure 8 presents an arc ramp when the packet rate increases from idle all the way to 100%. If we look at another data point at power consumption over variable packets size at line rate.

Figure 8. Packet Size/Traffic Load Power Consumption

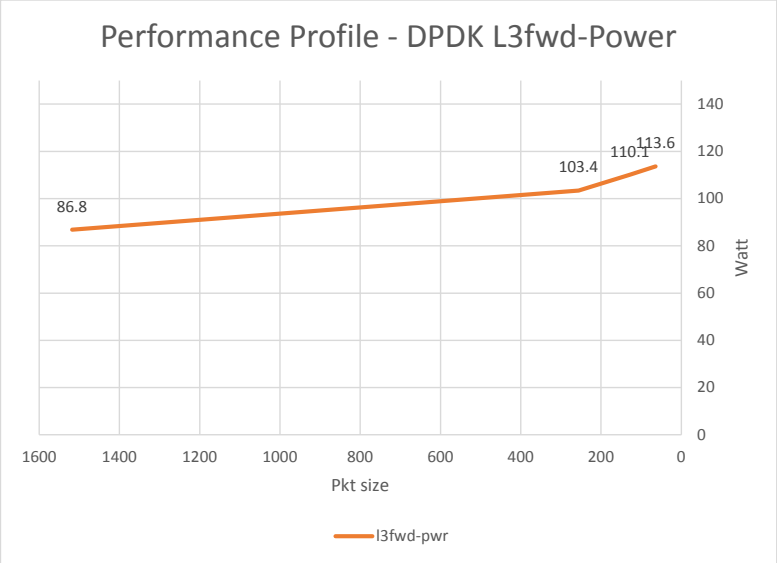
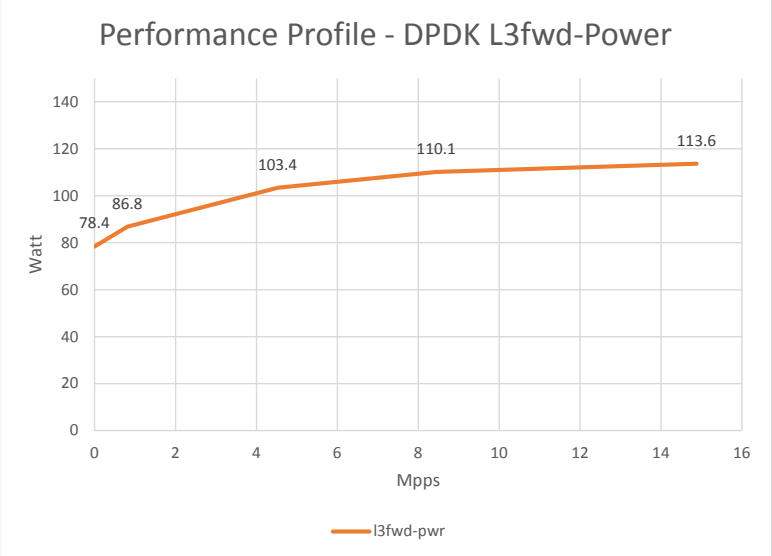


Figure 9. Packet Size (based on Mpps) Traffic Load Power Consumption

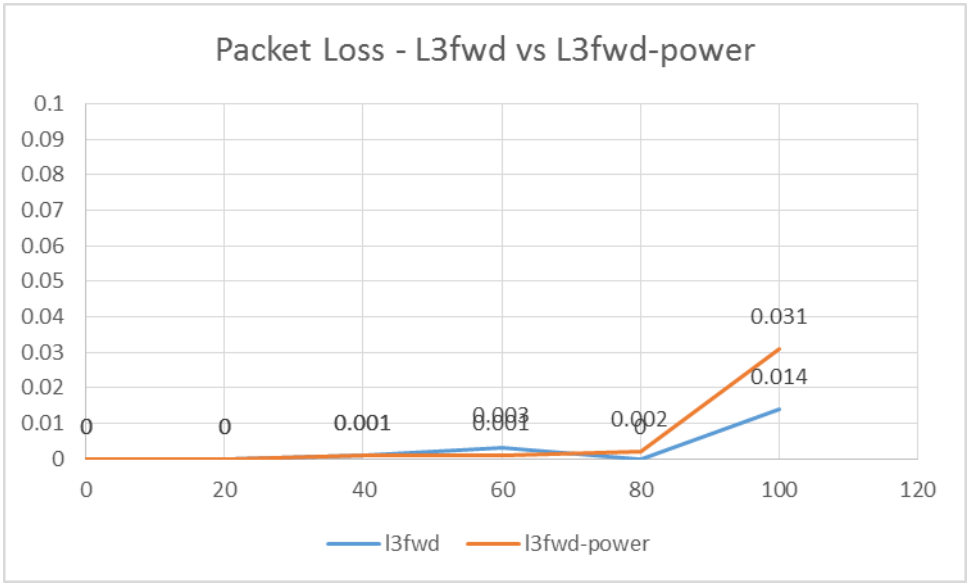


From Test Case 2, Figure 4 displays that as packet size decreases, the power consumption increase. This is caused by the workload load increasing as packet size decreases.

If the graph is plotted based on the packet rate for each packet size, it is evident that the similar arc graph with Test Case 1 is the result. It can be concluded that the power consumption of the sample application is directly related to the amount of packets need to be process, but not influenced by the data rate of network traffic.

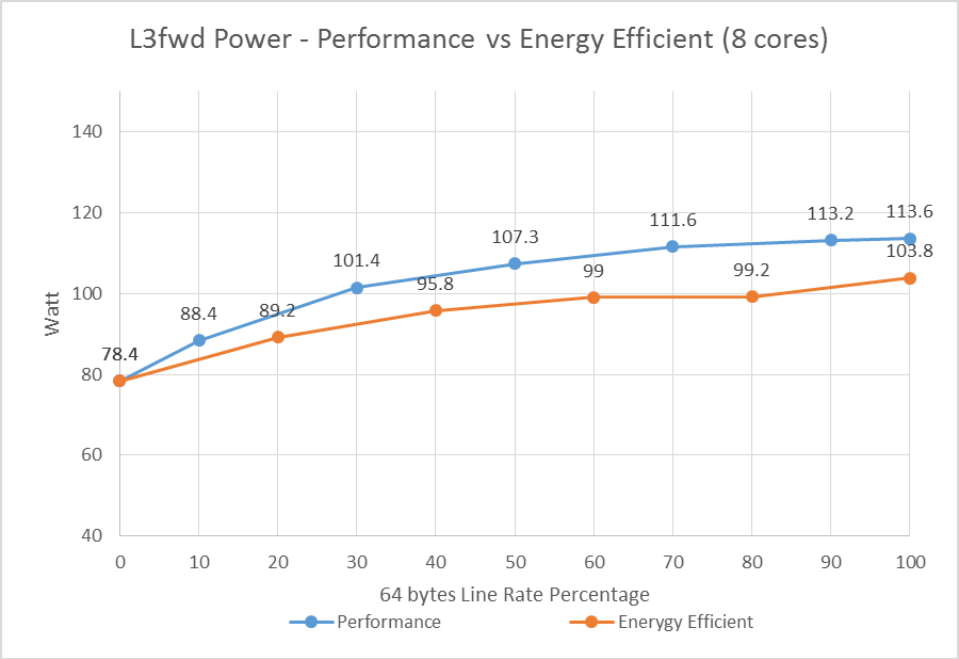
The explanation is for every packet that arrives into the system, the CPU is required to spend a certain amount of cycles to process it, regardless of the packet size. With larger packets, the bandwidth of the link (cable) limits the total number of packets coming in, which reduces the load to the system.

Figure 10. Packet Loss for Test Case 2



Looking at packet loss comparisons between L3fwd and L3fwd-power, there are no significant differences between both applications. The test was conducted based RFC2544, for a duration of 60 seconds.

Figure 11. Comparison Graph for Test Case 3: Different BIOS Power Management Profile for l3fwd-pwr



The idle powers are very close for both profiles; differences are not observable with the test duration being run.

A difference in power consumption could be seen though if the power profile changed from performance to energy efficient. At a workload between 0 to 60%, an energy efficient profile saved a few watts compared to the performance profile. Furthermore, the energy efficient profile starts to save more power (close to 10 watts) when the workload grew from 60% to 100%. The raw throughput is at roughly the same level for both profiles.

Even though the power savings are not huge when compared to switching from standard l3fwd to l3fwd-pwr, it still contributes to a maximum of 10 watts of saving.

§

3.0 Conclusion

The new Intel CPUs are getting more and more power efficient, but not at massive amounts. It is also worth noting that the CPU is but one piece of the puzzle in a network equipment.

Besides the hardware, software is also very important. It is proven that L3fwd-power sample application can reduce the power consumption of network appliances by more than 30%. Based on the performance degradation caused by the packet loss, the difference can be considered negligible. As L3fwd-power is a regular l3fwd, with additional features of reducing CPU P-states, and induce idle state allowing ACPI power management to put CPU into a deeper C-states.

Based on US EPA analysis, there is a vast difference between IT equipment and a data center server. Whereby IT equipment standard, US EPS assumes active power is 5% higher than idle power and traffic does not significantly impact power. This is similar with DPDK L3fwd.

For data center servers, idling is not common. Data centers push for high utilization to minimize capital expenditures for “peaking” capacity. Data centers minimize the number of extra, idle machines held in reserve.

With DPDK L3fwd-power, typical IT-equipment can be transformed to behave like a data center server in terms of power consumption. Based on a US EPA study, they assume that network equipment spends 25% of the time with high traffic (active state) and 75% of the time with low traffic (idle state). Hence, it shows that with a correct combination of software and hardware working hand in hand, a more environmentally friendly (power saving) data center or IT infrastructure can be achieved especially with Advantech* network appliances.

§