IMPACT ANALYSIS OF ENERGY PROPORTIONALITY MEASURES ON ENTERPRISE AND DATA CENTER NETWORKS

Aim:

1.To summarize the energy proportionality measures performed on energy proportional enterprise networks and data centers on networking devices such as switches.

2.To analyze the impact of energy proportional measures on networking devices based on the studies performed on the commercial IT enterprise networks in the light of issues such as switch's Buffer pressure, interface queue length and incast problem.

3.To analyze the extent of energy proportionality that can be achieved in networking devices of Data center networks using conventional energy proportional measures ; To speculate the possible new hardware aided approaches for switches which can be used with more confidence for rendering meaningful energy proportional savings or power saving systems; To study the impacts of those techniques on switch's Buffer Pressure, interface Queue length and the prevalent Incast of Data Center Networks which support Map Reduce technology.

Energy proportionality:

Barroso and Holze [1] found that the amount of energy spent is not proportional to the amount of energy supplied in google clusters. Numerous studies have shown that datacenter and enterprise networks rarely operate at full utilization, leading to a number of proposals for creating servers and networking devices energy proportional with respect to the computation they are performing. With the emergence of high bandwidth technologies in the order of hundreds of Gbps, the networking devices such as switches, routers needs to catch up with the speed and thus needs to consume more power. Moreover, being a shared resource, networking devices are always expected to be on and the utilization of the networking devices are not power proportional. Much of the time, they are under- utilized[2]. Thus, they are a good candidate for energy proportional computing.

Energy Proportional Measures: The approaches to arrive at energy proportional networked systems can be divided into four:

Profiling:

1.In this method Profiling the power usage of the devices and exploiting the idle periods found in the network traffic to save energy. It involves snooping the traffic in a networking device; looking for a pattern of usage in the energy; setting timers and planning the idle state measures, time period and reactivation methods(including obtaining tradeoff between the savings vs performance) ; operating the measures; optimizing the plans progressively as the traffic changes on a daily basis. Examples are the Mahadevan's Energy proportional enterprise networks[2], where they use a tool (Urja) to automate the above mentioned processes. Typical measures to save power in a networking devices(they are switched on even when not needed and are a good candidate to save energy)such as switches would include switching off the links or lower the data rates during idle hours(based on the assumption that lower data rate consumes lesser power). Another example is the Brandon Heller's energy saving Elastic tree topology for Data center networks[5].

Changing Network Topologies:

2.Satisfying the given workload using lesser number of devices through efficient topological constructs.

This approach involves understanding the capacities and configuration capabilities of individual devices in a cluster and the requirements of the system as a whole; modifying the underlying architecture to satisfy the system's requirements as and have lesser number of devices marginally greater than required. This is different from the former method which ensures energy proportionality for individual devices . Examples include the Dennis Abts and Micheal Marty's Flat Butterfly topology [3] for energy proportional data center networks . Here the system contains lesser number of devices, uses a rich network topology, and addresses the energy proportionality at system level. Another example is Elastic Tree. Here the aggregate traffic requirements for the data center is monitored and they networking devices are switched on based upon the obtained statistics to satisfy the load on the data centers.

3.**All-in Strategy**:

All-In Strategy (AIS). AIS uses all the nodes in the cluster to run a workload and then powers down the entire cluster to a low power state. There is no special hardware support suggested here which reduces the transition time. The benefits of energy savings amortizes the losses in the transition time.

4.**Power Nap:**

With the help of the hardware , the entire system transitions rapidly from high performance active states to near zero power idle states once task in hand is completed. Rather than requiring fine-grained power-performance states and complex load-proportional operation from each system component, PowerNap instead calls for minimizing idle power and transition time. The transition time between the hardware designed idle states to the active state is in terms of micro seconds suitable for highly volatile data center traffic such as Data Mining data centers.

Technique	**Traffic pattern/ Predictability**	**Networks applied to**	**Idle period/ Reactivation/ change of state time.**	**intuitions behind the approach**	**Energy savings**	**DrawBacks**
Power Profiling[1]	Clear and predictable	Enterprise Networks	rate adapting their ethernet channels takes 1–3 s.	Collect heuristics, predict pattern and switch off devices according to the pattern to save power.	the overall energy consumption can be reduced by up to 36%.	Uses coarse timing methods not suitable for highly volatile traffic of data centers.
Topology Revamp [2],[5]	Requires prediction of the bandwidth requirements of the plesiochronous Links	Data Centers	**Several ns to Several µs**. The locking process for receiving data at a different data rates is fast, 50ns–100ns for the typical to worst case	1.Clustered topology is better than hierarchical topology with lesser number of nodes and very high connectivity. 2.performs on the fly adaptive channel utilization using YARC routers [25]	1) Flattened butterfly uses 409,600 less watts than a fat tree with the same bisection bandwidth 2) 60% power savings compared to full utilization,	1. They still perform prediction of data center traffic for an epoch and rate adapt the link based on the measurement. 2. **The estimated power consumption when the chip is idle is 80W![25]**
	Predictable Aggregate load/traffic		Switching on a switch-30 secs/ powering on a port-1-2 secs.	Satisfy the predicted aggregate load using a subset of switches based Multi Commodity Flow problem.	energy savings ranging from 25-62%	Requires complete prior knowledge of incoming traffic to be fed into the MCF framework.
All-in Strategy [6]	Reacts to idle periods. Doesn't require a prediction/ Predetermined pattern.	Map Reduce Clusters	In several seconds.	1)avoids to choose the apt device to switch off to save power. 2)Save more power with ALL or nothing strategy Vs switching on/off . 3)has constant delay every time.	TeraSort on a relatively small 77GB dataset on a 24 node cluster at 33% utilization is 60% energy efficient with AIS Vs Unclustered .	1. High transitioning energy consumption 2.Huge transition time. (Not suitable of time sensitive applications.)
POWER NAP [14]	Reacts to idle periods. Doesn't require a prediction/	Data center	From µs - ms {depending upon the availability of	Hardware enabled near Zero power D2 states with faster transition time from	Power Nap and RAILS(power supply unit) redices average	Requires RAILS support for powering up power napping

	Predetermined pattern.		D2 states in devices}	the high performance active state exploits the micro delay present in highly unpredictable Data centers	server power consumption by 74%.	devices.

Enterprise Network:

An enterprise private network is a network build by an enterprise to interconnect various company sites, such as production sites, head offices, remote offices, shops, in order to share computer resources. In the Enterprise Networks, compared to other IT devices such as servers and laptops, energy efficiency of networking equipment has only recently received attention since networks, being a shared resource, are expected to be always on. Power consumed by the network is significant and growing even though the utilization remains . The estimated the annual electricity consumed by networking devices in the U.S. in the range of 6 - 20 Terra Watt hours [11]. According to "Enterprise Network Control and Management Traffic Flow Models in [12], The INMS SNMP agents collected data providing a 24-hour view of traffic sampled at 5-minute intervals shows that the network traffic is highest during normal working hours between 7am to 8pm with network utilization running at 5 megabits/second, (8% utilization) peeking around 11am-1pm at 12-megabits/second (27% utilization).The traffic pattern tends to concentrate more upon the peak hours of daily life and relaxes towards the night time. The point to note is that the utilization of the links is fairly low around 8-27%. Thus we have idle periods in the order of several seconds. Similarly, from the "Energy proportional enterprise networks " by Priya Mahadevan [2], there is a clean pattern of traffic flow in the enterprise network according to employer work duty. By profiling the link for the rate of flows, SNMP traces of data transfer, the patterns of traffic can be generated approximately. Given the traffic pattern and an optimizer which increases the accuracy of the pattern by continually gathering data, the load can be predicted for a certain particular period of time with buffer periods to avoid overflow in case of . When the traffic predicted is less than a threshold value, the link can be switched off to save power for a predicted amount of time.

This project studies the impact of performing the above mentioned techniques to save power in the light of buffer pressure, interface queue length of a switch or a router. We also discuss the impact of a rare (in enterprise networks)but possible phenomenon called Incast in a switch.

DATA CENTER NETWORK:

They are large clusters of tens to hundreds of thousands of servers in a farm providing services such as Search Engines, Email, Gaming, Advertising, Retail, Data Mining employed in university, huge organizations such as Facebook , Google and cloud computing clusters. According to "Cutting the Electric Bill for Internet-Scale Systems" [13], Google consumes more than $38M worth of electricity annually. But unfortunately, in a typical data centers the average server utilization is around 20-30% [1]. Data centers usually follow the Buffered Clos or the Fat tree topology, arranged as an hierarchy of layers of switches and routers namely core or the Top Of the Row switches, the aggregate or the intermediate layer and edge layers, which connect directly to the end systems. According to Benson's "Understanding Data Center Traffic Characteristics" , around 60% of the core links and the edge links are actively being used. Under utilization might imply a huge opportunity for saving energy by operating the ideal traffic engineering schemes across over utilized and under-utilized links. But the exact set of links unused are constantly changing[8]. Data center traffic is bursty and Moreover user demand varies rapidly and/or highly unpredictable. Typical idle periods lasts few seconds or less than one confounding the simple energy conservation approaches[14]. Partition-aggregate or Map-Reduce design is the foundation of many large scale distributed services.

We specifically deal with the Map reduce Data centers in order to study the impact of energy proportional measures in the light of buffer pressure and suggest a hardware based switches utilizing the PowerNap technique [14] for Data center Switches to save energy for the data centers with highly volatile traffic.

A NETWORKING SWITCH:

We consider shared memory switches as most of the commodity switches are shared memory switches. The shared-memory (SM) switch, consists of a single dual-ported memory shared by all input and output lines. Packets arriving on all input lines are multiplexed into a single stream that is fed to the common memory for storage; inside the memory, packets are organized into separate output queues, one for each output line. Simultaneously, an output stream of packets is formed by retrieving packets from the output queues sequentially, one per queue; the output stream is then demultiplexed, and packets are transmitted on the output lines [16]. Each of the external interface has its own output queue filled with data ready to delivered. Buffer allocation determines how the total buffer space (memory) will be used by individual output ports of the switch. There are two basic buffer allocation policies: (i) Complete Partitioning and (ii) Complete Sharing. In the complete partitioning (CP) scheme, the entire buffer space is permanently partitioned among the N servers. The sum of the individual port buffer allocations is equal to the total memory M. In the Complete Sharing policy, an arriving packet is accepted if any space is available in the switch memory, independent of the server to which the packet is directed. In other words, individual buffer allocations equal the total memory space. The switch has N output ports, and a total buffer space of M. A first-in-first-out (FIFO) queue is allocated to each output port, denoted by kj• Packet losses occur when a packet arrives at a switching node and finds the interface queue full.[17]. We assume that a common buffer for the entire switch and Shared buffers are placed in the output ports[18]. The complete sharing method results in delay due to path difference between the best path and the available path for the cells to be sent. Naturally, the best of the two methods is perform the sharing only when the assigned interface is full called the Dynamic Partial Sharing .

$$\textit{Complete Partitioning}: \quad \sum_{i=1}^{N} ki = M$$

$$\textit{Complete Sharing}: \quad ki = M, i = 1 \ldots N$$

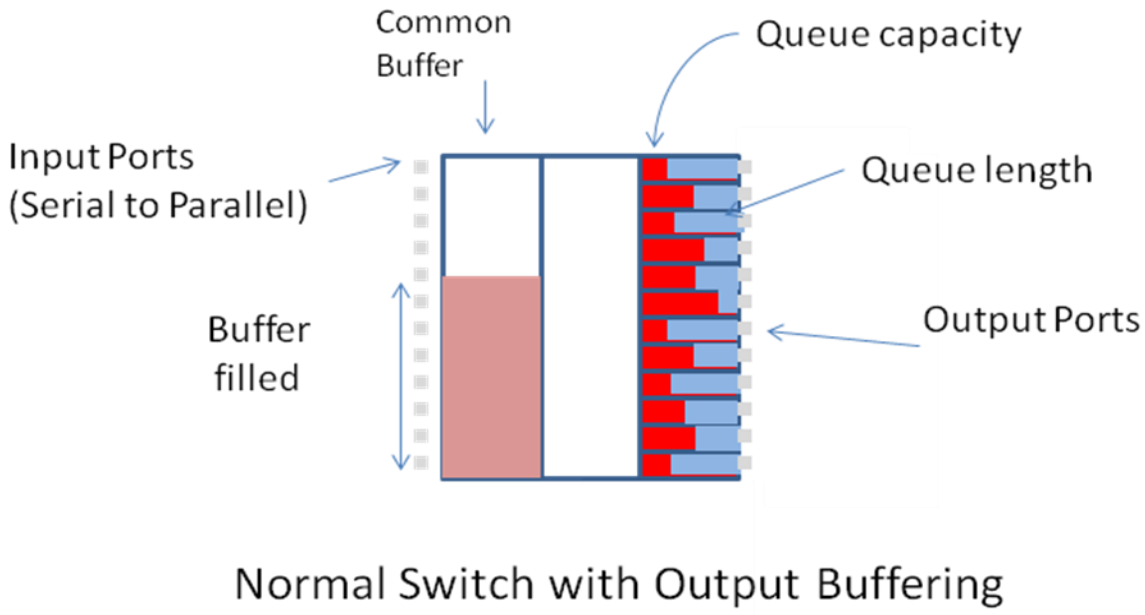
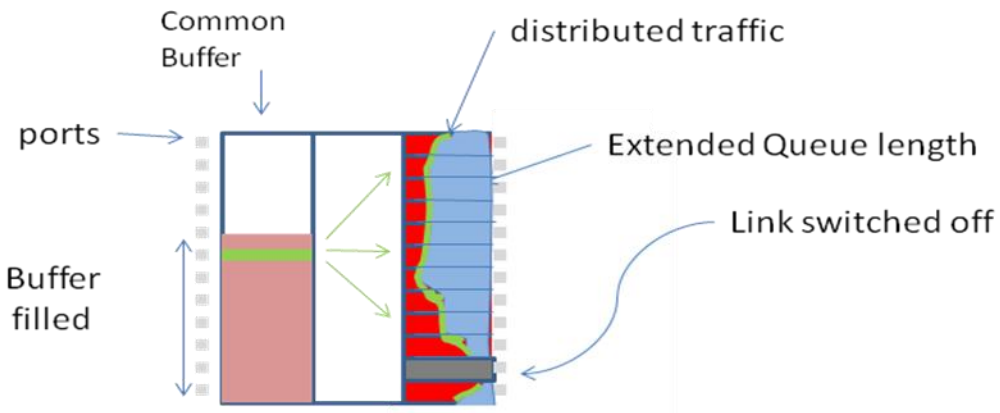


Figure 1

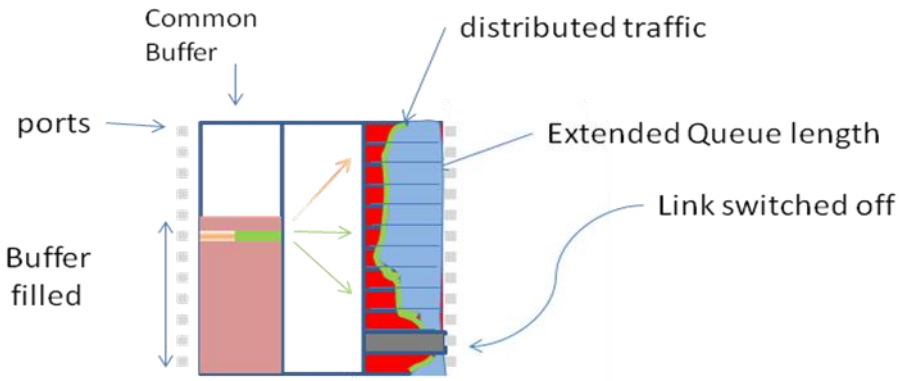
IMPACT OF ENERGY PROPORTIONAL MEASURES ON ENTERPRISE SYSTEMS:

Switching Off Links:

We assume shared switches in Enterprise networks follows Complete Sharing buffer allocation once the addressed interface queues are full called Partial sharing method. In an enterprise network switches we have a definite pattern diurnally varying based on the usage profiles of the employees over the systems in the corporate sector [12]. Once the traffic pattern is obtained by means of suitable prediction method. By following the Profiling method described earlier the energy is saved by means of switching off the links. When such a measure is taken, in addition to saving energy there are certain overheads for the switch. It is possible that we do not find the zero load link to switch the link off in the enterprise network[12]. We can still take the advantage upon Complete Sharing buffer allocation followed in the switches and decide upon the lowest loaded link to be the target link to switch off for the quantum of time to save energy. When a non zero-load link is switched off , the switch looses that link to distribute data load of the target link. This results in building up of the cells in common buffer called the buffer pressure and by the virtue of complete sharing buffer allocation on the interface queues[17], the traffic can be reduced to save loss of cells. The result is a marginal increase in the length of the output queue in each of the external links of the switch . Here, depending upon the Virtual LANs configured in the switch we will only have a subset of the remaining links to share the load of the 'dead' link. The transition time for this process is in the order of several seconds . It results in naturally more throughput delay as now the queue is slightly is larger for the data to flow in the switches. It also has the side effects for huge delay for packets from delay sensitive applications . Simple conceptual diagram below illustrates the method of switching off the links to save power. Depending upon the amount of idle traffic more than one link could be switched off to save power. In that case the amount of delay experienced by an delay sensitive application will be magnified as the number of links available to reduce the load on the buffer is reduced. Thus a local mechanism which follows energy proportionality methods according to ONLY the specific link in hand will result in huge packet loss . Several mechanism follow such a link specific method to save power[1]. They would drain the switch's capacity when the load is less. The loss can be reduced only when the transition time is significantly low. But for the switches during switching off the links the transition time is in the order of seconds and additionally they follow link specific heuristic to save power. This would result in loss of packets.



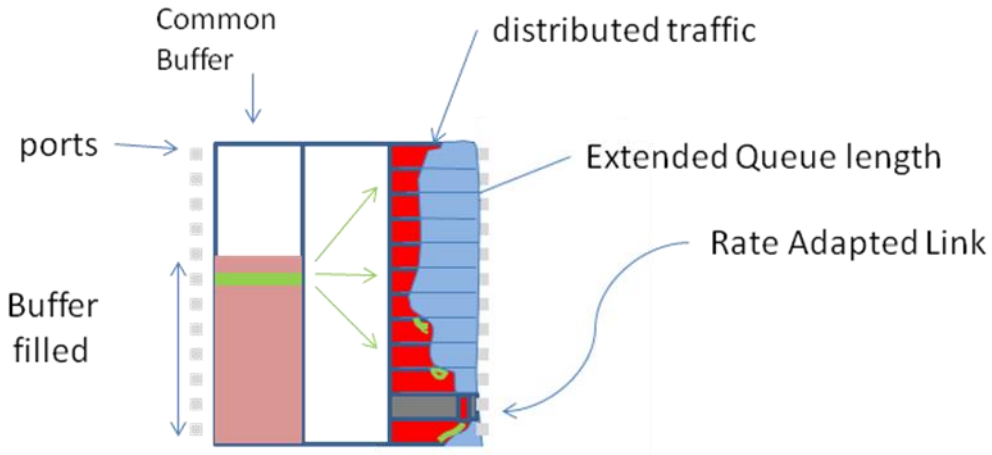
Energy proportional Enterprise Link Switch



Energy proportional Enterprise Link Switch with a short urgent msg.

Line Rate Adaption:

When we follow the rate adapting the links in the Enterprise networks, The links which are operating in the order of Gbps are powered down to operate in Mbps to save power when the traffic is in the link around 10-25%. This is easily achieved in the enterprise networks as the utilization of the links is low in the order of 8-27%. It adapts to the incoming buffer rates. In this method there is significant avoiding of loss of cells as the interface queue being rate adapted still can allow delay sensitive packets to be queued efficiently. The only drawbacks of this approach are that the down time for rate adaption is found to be 2-3s .The approach is perfectly fine for enterprise networks where the idle periods are in minutes or hours [19],[2]. There needs to be some slack capacity added the bandwidth of the energy proportional link to reduce the loss of packets due to incorrect prediction or sudden surge in the traffic load on the switches. Traffic cannot always be predicted perfectly. This situation would be smooth when the number of links rate adapted is few. But as we increase the number of links rate adapted the total buffer pressure on the common buffer is increased significantly.



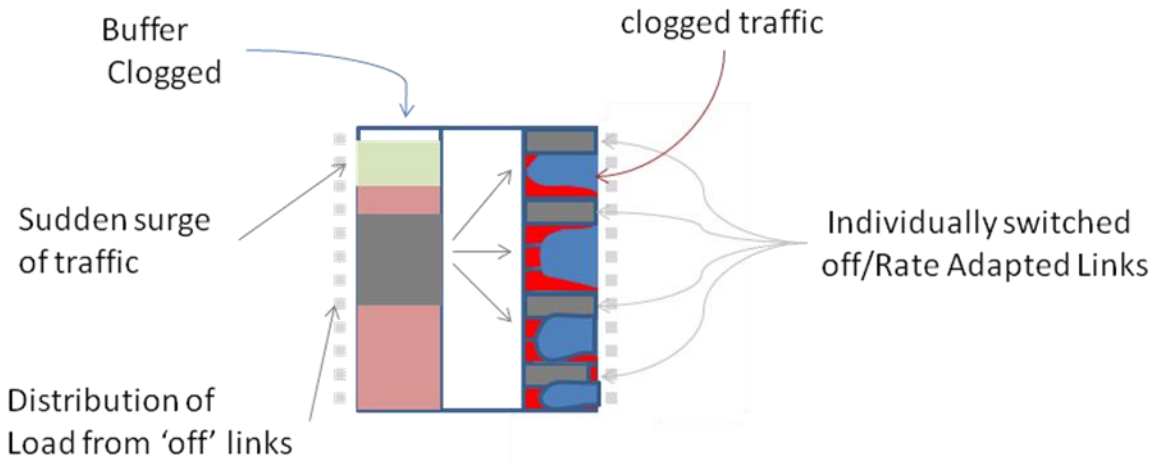
Energy proportional Enterprise Link Switch

The Buffer pressure:

The incoming cells are stored in the common buffer of the switch till they get transferred to the output interface queues to be delivered. when the links are switched off the cells stay in the common buffer increasing the 'pressure' on the buffer. The energy proportionality measures do not consider the amount of occupancy of the common buffer when they switch off the interface queues. For Complete partitioning switches, the energy proportional measures would lead to heavy packet loss as strict segregation of the cells to follow the single output port would result in detrimental loss. For complete sharing , since the queues are shared we would see lesser impact of switching off the link as the load is shared by all the interface queues. The common buffer will be loaded more during the energy proportionality measures when the remaining interface queues are full and thus have high usability than the complete partitioning method.

BUFFER BURST:

Algorithms used to save power by means profiling uses link specific measures to save power[2]. By link specific' I mean to say is that the heuristic to switch off the link/ rate adaption is done only based on the current traffic in the link under consideration. The problem with this approach is that when the traffic is lesser for the links, and many links will be found to have less traffic flowing through them. Now the aggressive method to save power will either switch off/rate adapt multiple links as the traffic observed will be lesser than the threshold during idle periods in the enterprise networks. Thus the buffer pressure will rise significantly to distribute more packets when the prediction fails or due to sudden surge of traffic in the enterprise networks. The problem is intensified by the delay produced due to transition time (2-3s for rate adaption and higher delay for the switching on the link) from the low power state to the active state by the link. The result is significant loss of packets if multiple links are switched off in the switch. we call it the BUFFER BURST. This would result in catastrophic delay for the delay sensitive applications .



BUFFER BURST with sudden surge in the traffic

Energy Proportionality on a single switch link:

Let R_i, be the link chosen to be switched off for *t* seconds, where *i* can be one of 1,2,3,...N for N is the total number of links available in the switch to dispatch the traffic. We assume that the capacity of each output link is same. The capacity(bandwidth) of each link be μ bits per second. The threshold traffic below which the link can be chosen to be switched off be μ_T. μ_T is typically given by 10-15%. The current traffic in the Link R_i be given by μ_{Ci}. When a link is switched off, the bandwidth μ_{Ci} is shared equally between the remaining links in the switch according to the Complete sharing mechanism. The Length of Queue for link L_i can be denoted by Q_{Li} . ΔQ_{Lx} be the change in the Queue length for each of the links in the switch due to the removal of a link L_i. where $x \in (1,2,3..N), x \neq i$. The increase in the queue length of each remaining link, given by:

$$\text{Increase in Queue Length } (\Delta Q_{Lx}) = \frac{Bci}{N-1} \qquad (1)$$

If we have L out of N links being used for VLAN, then the number of links available to distribute the traffic is given by N-1-L. Thus the average increase in the queue length can be given by

$$\text{Increase in Queue Length } (\Delta Q_{Lx}) = \frac{Bci}{N-L-1} \qquad (2)$$

When a link L_i is switched off, the cells to be transferred via that link gets accumulated in the common buffer waiting to be distributed to the remaining links. The maximum increase in the buffer pressure P when we remove the link L_i for time T_i can be denoted by ΔP_{Li}.

$$\text{Increase in Buffer pressure } (\Delta P_{Li}) = Bci \times T_{Li} \qquad (3)$$

BUFFER BURST:

Let $R_{i,..}$ R_j be the links chosen to be switched off for corresponding $t_{i...}$ t_j seconds, where $1<= i<=j <=N$ can be one of 1,2,3,...N for N is the total number of links available in the switch to dispatch the traffic. The capacity(bandwidth) of each link be μ bits per second. The threshold traffic below which the link can be chosen to be switched off be μ_T. The current traffic in the Link R_i be given by μ_{Ci}. When the links are switched off, the bandwidth $\mu_{Ci,...}$ μ_{Cj} are shared equally between the remaining links in the switch according to the Complete sharing mechanism. The Length of Queue for link L_i can be denoted by Q_{Li} . ΔQ_{Lx} be the change in the Queue length for each of the links in the switch due to the removal of a link L_i. where $x \in (1,2,3..N), x \neq i.$ The increase in the queue length of each remaining link, given by from (2):

$$\text{Increase in Queue Length } (\Delta Q_{Lx}) = \frac{1}{N-1-L} \sum_{k=i}^{j} \mu ck$$

When the links R_i ... R_j are switched off, the cells to be transferred via that link gets accumulated in the common buffer waiting for the availability of remaining links to be distributed. The maximum increase in the buffer pressure P when we remove the links $R_{i...}$ R_j at time T can be denoted by ΔP_{Li}. This occurs when the remaining buffers are filled or involved in other VLAN connections.

$$\text{Increase in Buffer pressure } (\Delta P_R) = \sum_{k=i}^{j} \mu_{Ck} \times T_{Lk}$$

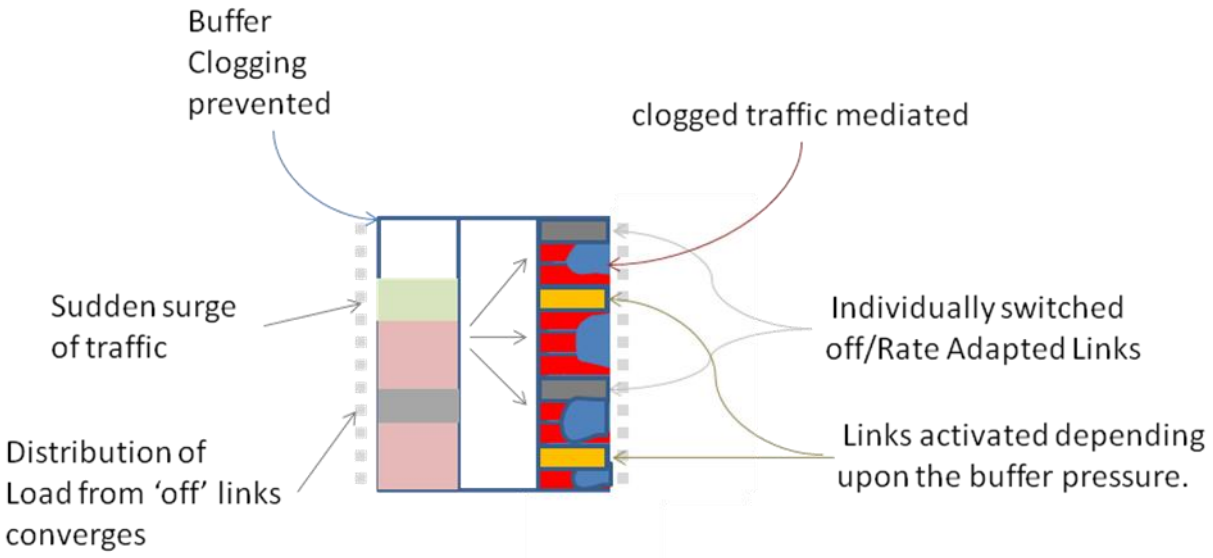
It should be noted that the individual links will put to sleep while such a process is happening. Thus the value of μ_{Ck} is around 10-15% of the bandwidth of the link R_k. When one of the link which was sleeping comes to active state then the traffic is unloaded on to that link.

SOLUTION TO BUFFER BURST:

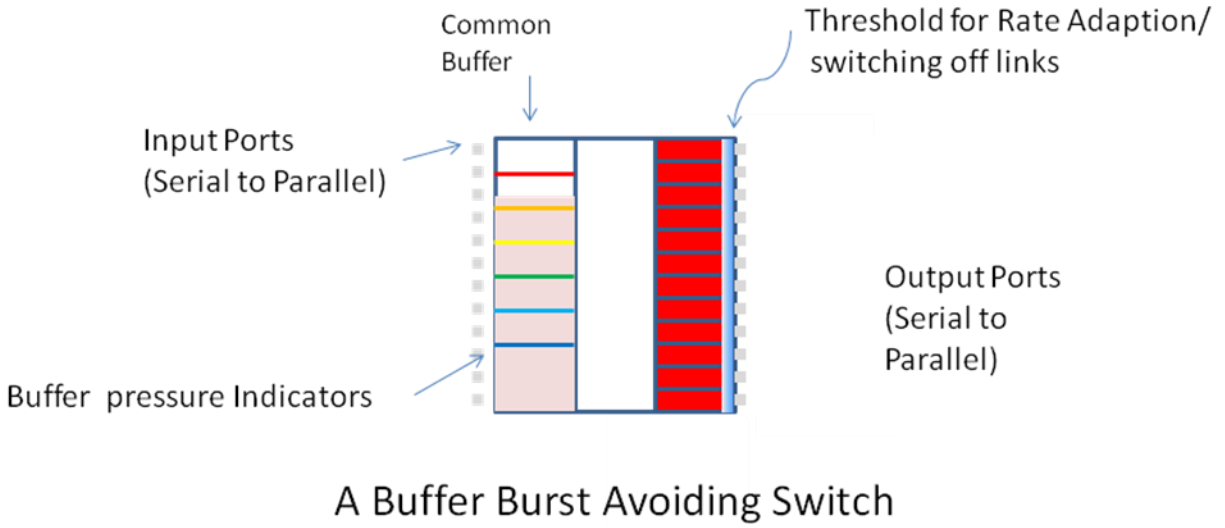
The solution to this problem involves use of Efficient Hardware to transition from low power state to the active state. Or we can mitigate this problem by means of software methods. In addition to this the heuristics used to switch off the link or rate adapt the link , the buffer pressure should also take into consideration. The approach to solve the Buffer Burst is to have threshold on the Common buffer pressure in addition to the having a threshold on the individual links when making a decision on switching off or line rate adaption for a switch link. If the buffer has a thresholds CBTs(Common Buffer Thresholds) depending on the load to pronounce the maximum number of links which can be switched off to save power based on the link rate to avoid Buffer burst . In addition to the predicted traffic towards the link for the particular period, we can have the occupancy of the common buffer to decide the number of links that can be switched off to save power by means of profiling. We can have indicators to buffer occupancy to decide the number of switch links that can be switched off or rate adapted to save energy. Or We can also have percentage of total number of links that can be switched directly related to the percentage of free space in the common buffer. We thus formulate the maximum number number of Links that can be switched off to save power (L_{MAX}) to the maximum number of links available in the switch N. as the following.

$$L_{MAX} = N(B_{FREE}/B)$$

Here B_{FREE} is the amount of free space in the Common Buffer. While B is the maximum capacity of the common Buffer. This would mean that we need to monitor the buffer pressure in addition to monitoring the individual link traffic to decide upon the energy proportionality measures. In a typical switch when the common buffer is getting filled up with the incoming packets the indicators get marked from the monitors to check the buffer pressure. The Lmax value is a function of the buffer pressure. It is directly proportional to the free space available in the buffer signifying that the more the free space in the buffer to more the links can follow the individual link optimization can be done and vice-versa.



BUFFER BURST MEDIATED



A Buffer Burst Avoiding Switch

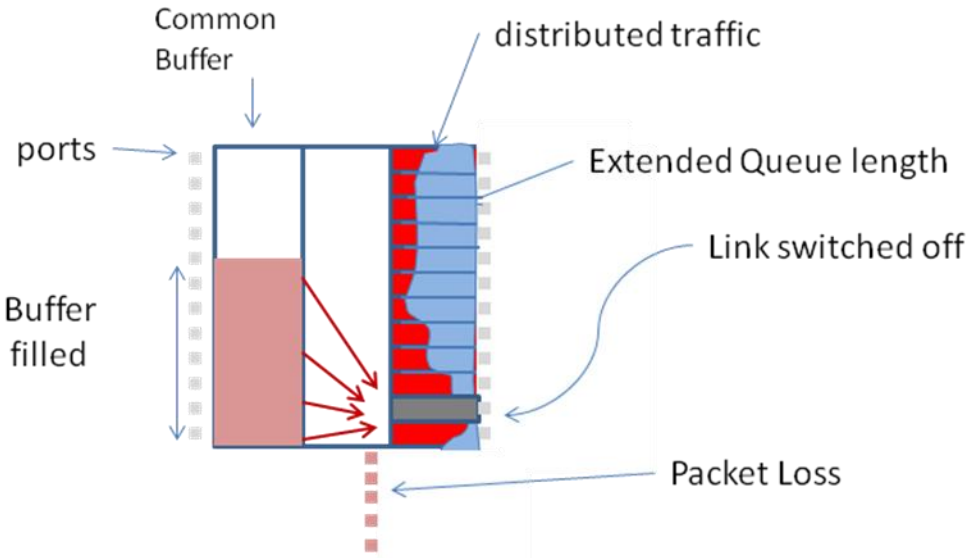
INFINIBAND AND YARC ROUTERS WITH LINK RATE ADAPTION: An interesting observation is made in the paper "Energy proportional data centers"[5] about the link rate adaption that could be performed using YARC/Infiniband Routers[25]. InfiniBand and YARC allow links to be configured for a specified speed and width, with the reactivation time of the link can vary from several nanoseconds to several microseconds. For example, when the link rate changes by 1x, 2x, and 4x (e.g., InfiniBand's SDR, DDR and QDR modes), the chip simply changes the receiving Clock Data Recovery (CDR) bandwidth and re-locks the CDR. Since most SerDes today use digital CDR at the receive path the locking process for receiving data at a different data rates is fast,**50ns–100ns** for the typical to worst case. Now with the Infiniband Routers the impact of the Buffer burst can be tremendously reduced. This is highly suitable for volatile traffic of data center networks.

INCAST IN ENTERPRISE NETWORK:

The event of sudden movement of flows from many of the input links to a single exit link in a switch results in overflowing the queue of the exit interface resulting in catastrophically huge loss of packets called incast. TCP incast congestion happens in high-bandwidth and low latency networks, when multiple synchronized servers send data to a same receiver in parallel [15]. For many important data center applications such as MapReduce[5] and Search, this many-to-one traffic pattern is common. The occurrence of incast in an enterprise network is rare. This is because the chances of many to one transaction are rare in the networks. Many to one transactions need to be time synchronized to arrive at same time at the switch to result in the incast congestion. The transactions in Enterprise are not as synchronized as Map reduce applications (they have many to one flow during the reduce stage) to end up in Incast congestions.

If it occurs in the enterprise switches, it is very difficult to track the links and mark them as incast vulnerable. The probability of finding the incast is higher for switches which are facing towards the servers. Intuition behind such an assumption is that the requests from numerous client systems would converge to one or few servers and thus would end up in a possible incast(unlikely though). But if such an incast occurs there would be loss of packets due to excess when the switch is least prepared to handle it, then the results is a huge loss of packets. An extremely rare event of incast might directly attack the inactive link switched off to save power.

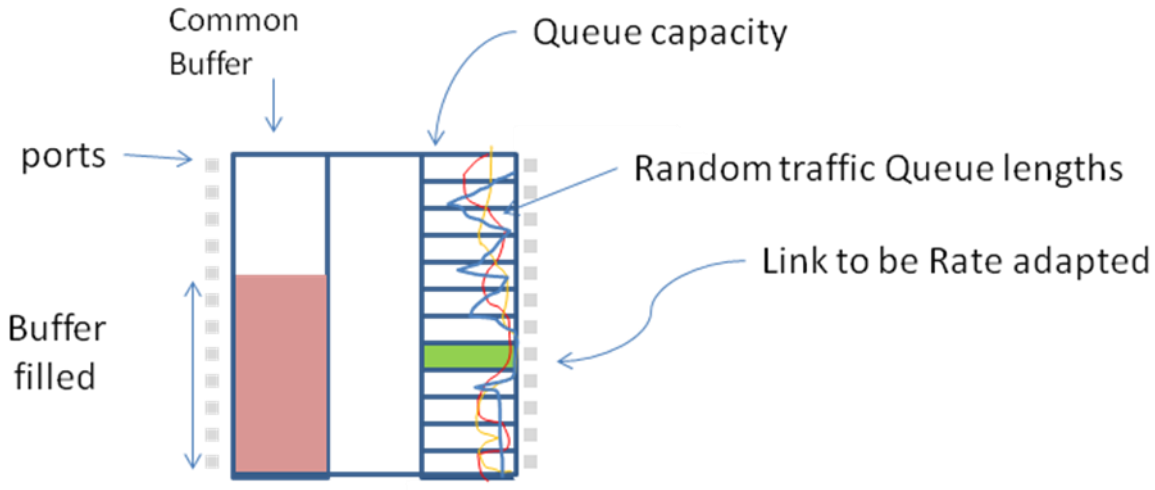
We also have a extremely rare possibility of BUFFER BURST AND INCAST completely choking the switch. But the probability is far too less. We have infiniband link rate adaption to the rescue though. The scope of which are beyond this project.



INCAST targets the dead link

DATA CENTER TRAFFIC:

Data center traffic is highly volatile. An analogy can be given for energy proportionality between the enterprise networks and Data centers. Wings of the butterfly Vs Wings of a bee. We can see the top of the wings of a butterfly with our normal eye clearly as the rate of flapping is slow but not the top of wings of a bee as it flaps extremely fast . It doesn't mean that the bee never keeps its wings flat for a while. It is just that our eye cannot capture it. Similar rule applies for the observable and tangible enterprise network traffic idle periods Vs the idle periods found in the data center traffic. The observations made in the data center traffic and the energy proportional methods applicable can be tabulated as follows. The method targeted is individual link oriented profiling .



Data Center Network Switch

Paper	**Application**	**# of DCs**	**Traffic Nature , Patterns and Predictability**	**Idle Periods and their nature**	**need for faster transition to active state from sleep state**	**Energy proportional Techniques possible And why?**
VL2[7]	Data Mining (Map-Reduce)	1	Highly tremendous. cycling among 50-60 different patterns during a day and spending less than 100 s in each pattern at the 60th percentile	No mention in the paper	10ms as the maximum acceptable response time	Power Nap NICs; Infiniband links. For other techniques, The delay produced are in the order of s. 2. Other techniques Involves heavy traffic predictions ,patterns, repository (elastic topology revamp)
Data Center Traffic [8]	Cloud DCs Map Reduce Web Services	19	evidence of ON-OFF traffic patterns between the packets.	unused links (40%). 80% of the unused links are idle for 0.002% of the 10 days or 30 minutes.	Inter-arrival times for packets according to the clock granularity of 10 µs.	Infiniband link rate adaption can be used. the reactivation time is in the order of several ns- 10 µs[5]
Nature of Data center traffic [15]	Cloud DC Map Reduce	1	work-seek-bandwidth and Scatter-Gather patterns in datacenter trac as exchanged between server pairs in a representative 10s period	The inter- arrivals at both servers and top-of-rack switches have pronounced periodic modes spaced apart by roughly 15ms	86% of the links congested for >= 10s.15% of the links congested for >=100s.	Power Nap for switching off the devices Vs Rate adaption of Inifiniband links.
Elastic Tree[5]	Any DC Map Reduce Web Service	1	Aggregate traffic peaks during the day and falls at night providing a diurnal aggregate pattern.	They have found aggregate low periods in the order of 100 minutes.	Irregular spikes in router port utilization	Aggregate topology revamp{if traffic matrix is available apriori}, Power Nap NIC, Infiniband/YARC links.
Traffic Of DCs In wild[4]	Cloud Private Net Universities, Map Reduce Web service Distributed F'S	10	the packet arrivals exhibit an ON/OFF pattern at both 15ms and 100ms granularities.	an ON/OFF pattern 15ms and 100ms granularities.	new flows can arrive within rapid succession (10µs) of each other, resulting in high instantaneous arrival rates;	Power Nap NIC; Infiniband links
Fine grained TE in DCs [21]	Cloud MapReduce	1	approximately 35%or 0.35 of the total traffic exchanged between pairs of ToR remains predictable. it is predictable for 2 seconds	Not mentioned in the paper	TE must react in under 2 seconds.	PowerNap NIC and Infiniband Links, For other techniques, The delay produced are in the order of s. 2. Involves heavy traffic predictions ,patterns, repository

						and computation.

Clearly the methods to save energy by means of profiling requires the pattern to be concrete; the delay characteristics to be in the order of seconds.(to perform Line rate adaption or switching off the switches). There needs to be diurnal patterns to apply the Aggregate profiling[5]. With the traffic characteristics showing phenomenal volatility in most of the data centers mentioned above , or data center traffic showing tangible patterns but still lacks the precision to saving energy during the huge number of micro-idle periods found in them, the best method to save energy would Power Nap technique[14] when the order of the granularity is in hundreds of µs. For the idle periods lesser than the 10 µs we can follow the Infiniband/YARC routers whose rate adaption requires transition time in the order of Nano seconds!

POWER NAP: With PowerNap, the entire system is transition rapidly between a high-performance active state and a minimal-power nap state in response to instantaneous load. Rather than requiring components that provide fine-grain power-performance trade-offs, PowerNap simplifies the system designer's task to focus on two optimization goals:

(1) optimizing energy efficiency while napping, and

(2) minimizing transition time into and out of the low-power nap state.

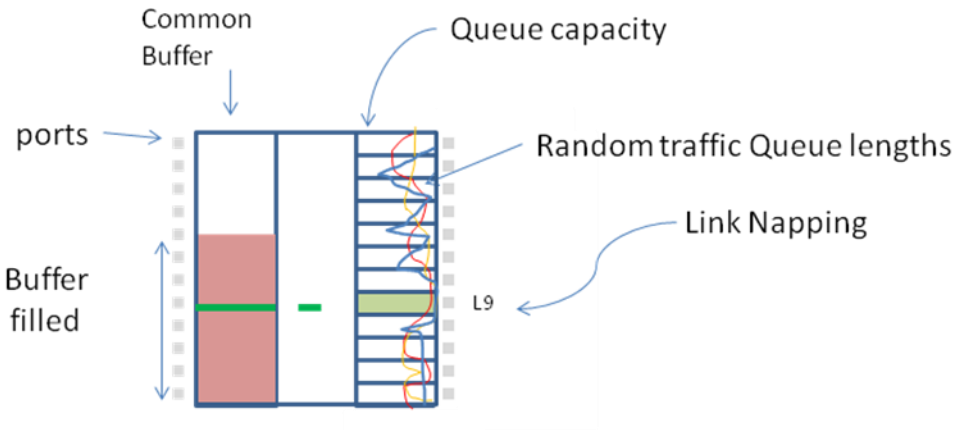
There are no patterns to learn or computation to predict the pattern or apply the pattern temporarily .

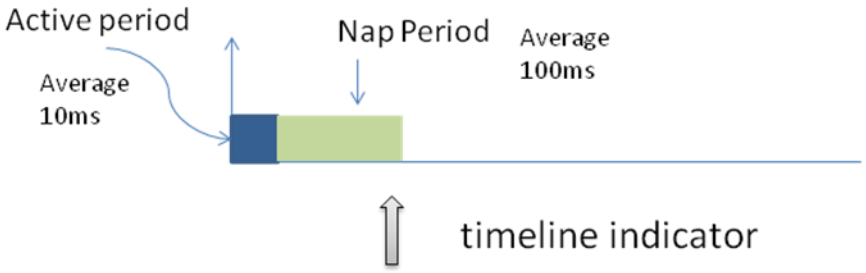
POWER NIC: The key responsibility PowerNap demands of the network interface card (NIC) is to wake the system upon arrival of a packet. Existing NICs already provide support for Wake-on-LAN to perform this function. Current implementations of Wake-on-LAN provide a mode to wake on any physical activity. This mode forms a basis for PowerNap support. **Current NICs consume only 400mW** while in this mode[14].

POWER LINKS: From the YARC/Infiniband router[25] the Link rate adaption can be enabled to save power. The rate adaption is in the order of several nano seconds[5]. But still the amount of idle power consumed is high as 80 W which is very high when compared to that of the Power NIC. Thus the idea is to combine both the NIC and Infiniband/YARC Links to have a POWER SWITCH which has **rate adaption latency of several ns-10 µs at the worst case [5]** and consume 80W at idle state and has **switch off latency of 200 µs and consumes as low as 400mW** power during the Zero State.

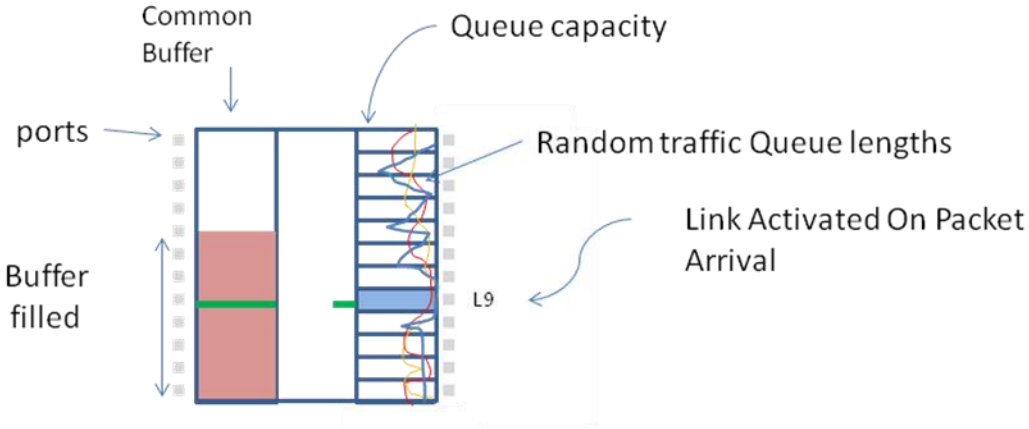
POWER SWITCH: A switch which has the Power NIC links in addition to the Power Links supported by infiniband will solve the granularity problem we have been in the other methods to save energy. As the LAN card is activated on the arrival of the packet, the switch can switch off aggressively. By aggressively, I mean to say that there is no need to snoop the link for the pattern present in the traffic or predict the future traffic based on the past repositories. The links can be switched off to the zero state, the moment the interface queue is emptied by the link. This holds good for the Link rate adaption of the Power Links [25]. Thus we have two choices to save power now. We can follow Line rate adaption using the infiniband links or switch of the links using power Nap. Either of the techniques can be followed depending upon the traffic in the data centers. On the arrival of the packet from the common buffer to the exit interfaces the link wakes up with least amount of Transition time provided by the specially designed D2 states of power switch.[14]. The LAN Cards given in [23] are PCI 1.1 power management specification complaint. According to the PCI 1.1 Specification[22]: **minimum recovery time requirement of 200 µs between when a function is programmed from D2 to D0.** Now this gives significant leverage to transition the links to the low power state D2 at minute granularities in time when there is no task in hand or there is no packet in the interface queue to send in same line as the Power Nap technique or we can perform the line rate adaption to reduce the amount of power consumed and be ready to face adverse problems such as incast for incast prone links.

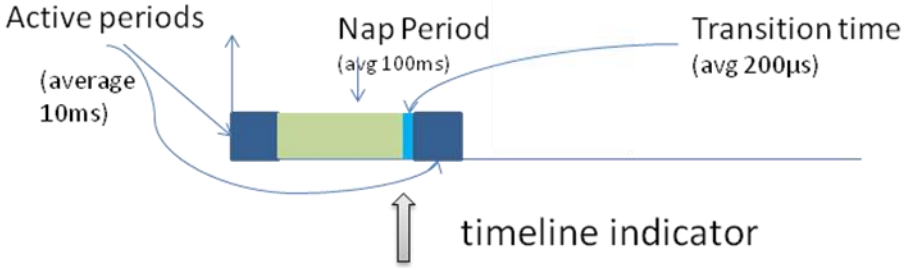
Power Napping Data Center Switch





Power Napping Data Center Switch





POWER SWITCH OFF : The downtime provided according to the PCI 1.1 specification for a transition between D3 Off state to D0 is around 10ms. This can be used in enterprise networks and Data centers whose traffic is in the granularities of seconds(rare) to save significant amount of power. Thus we also reduce the pattern matching computation for enterprise networks though this involves replacing all the available switches with the power switch. It is a tradeoff between cost of buying these energy proportional switches and the energy saved by means of new power switches.

DRAWBACKS WITH POWER NAP:

1. They require specifically designed RAILS to provide micro granularity in power management.[14]

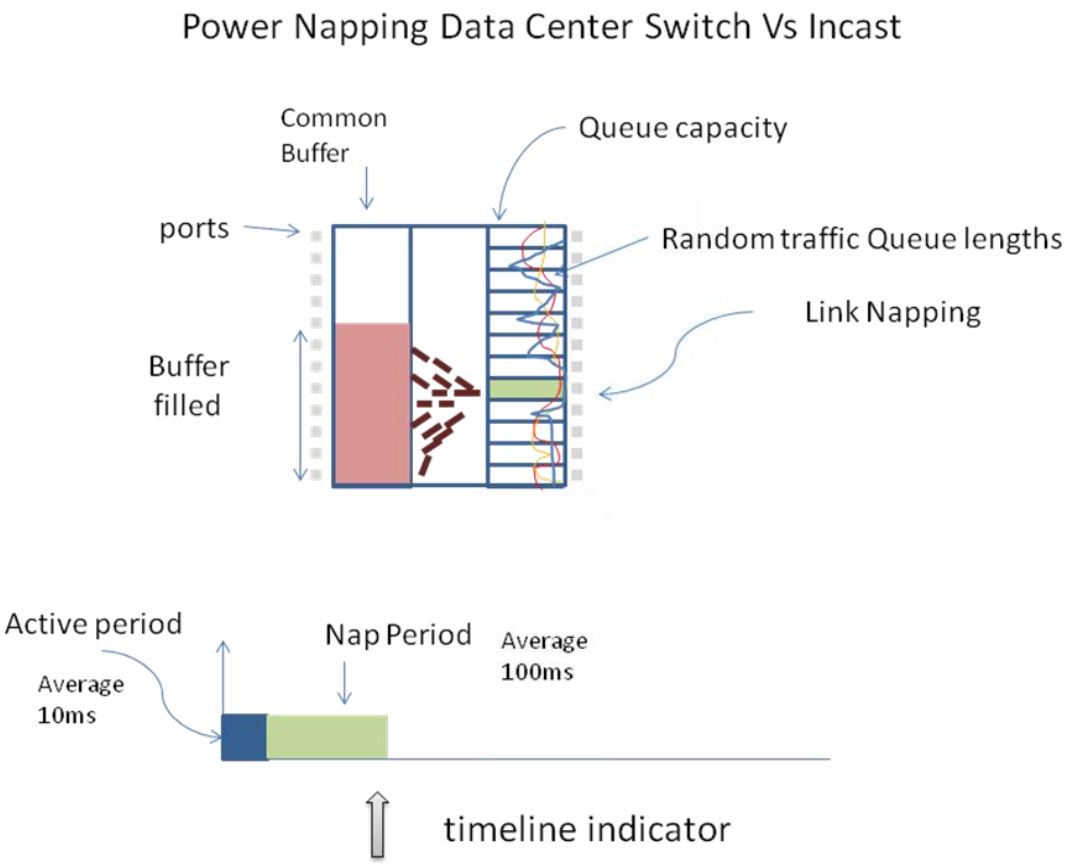
2. The method is aggressive.i.e. the try to save energy whenever the link is free. This is good for saving maximum energy in data centers. But this could also result in certain side effects due to incasts leading to a phenomenon, which I call as 'Cluster Cast' in Map Reduce Data centers. The probability of Cluster cast is less though. The problem of cluster cast can be alleviated by choosing the link rate adaption for incast prone links(also explained below).

MAP REDUCE:

MapReduce includes the Map and Reduce two program operations. MapReduce makes a master server control many data servers to execute concrete Map tasks and Reduce tasks. Map operation is applied for data classification and preparing intermediate data for Reduce operation. Reduce operation is applied for merging the intermediate data according to defined programs and preparing data for the new Map operation. MapReduce can process terabytes of data through many iterative Map and Reduce steps. During the reduction step, the completed operations return to the server. If the jobs are synchronized the incast phenomenon might take place in the switch connecting those multiple intermediate client machine to the single intermediate server machine. This is the main cause for incast in data centers.[24].

Power Nap Vs Dead Link during Incast:

When the link is idle in an enterprise network with zero-power-D2, the amount of loss of packets were phenomenally huge due to the inability of the link to respond to the heavy incast traffic load within short span of time. On the other hand the Power Nap switches can transition from near-zero D2 state to active D0 within 200µs. So during incast, the amount of packets are significantly reduced by the power nap switches. But Still there would be a very minute delay during incast in power nap switches, when the link wakes up to see a flood of incast packets while making a transition.



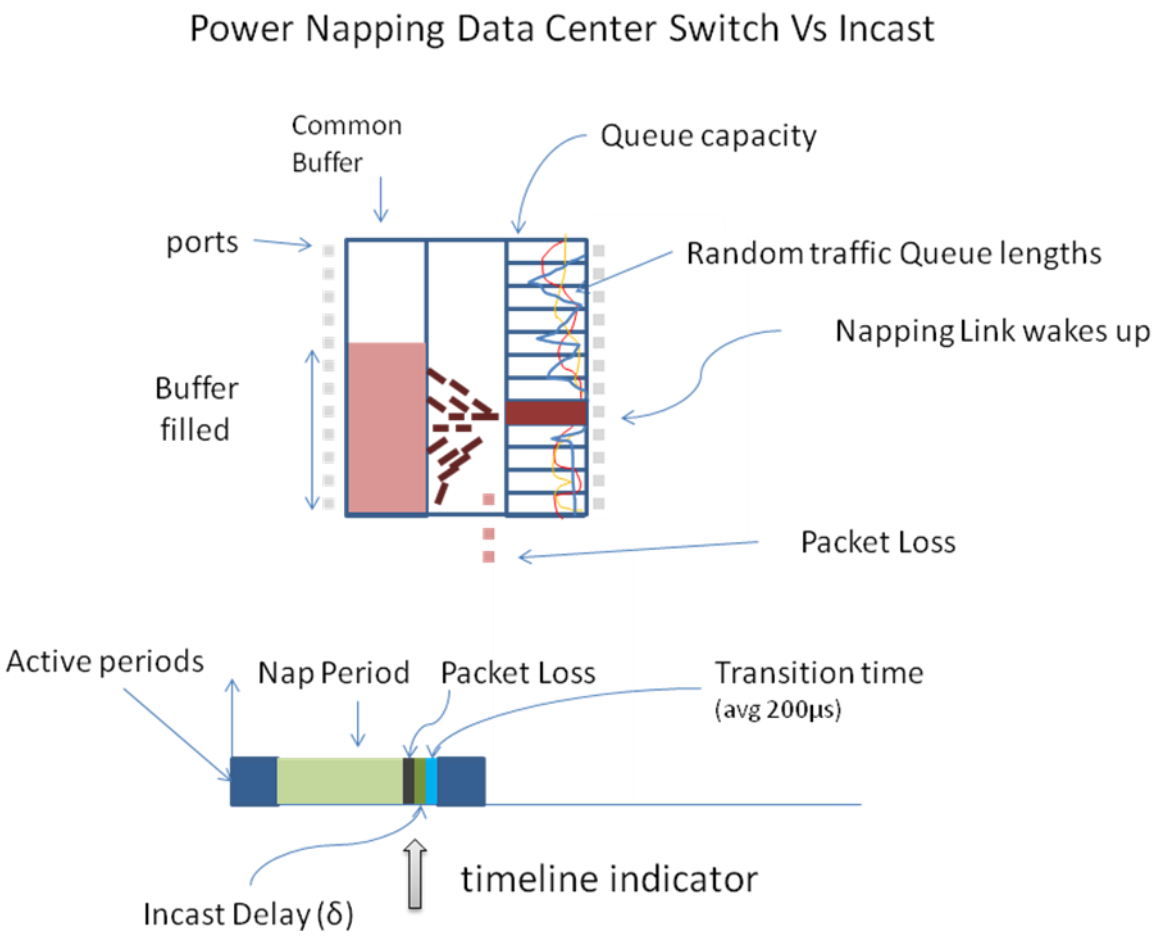
This delay corresponds to pressure on the interface link when handling a phenomenally huge number of incast congestion packets versus handling a few packets when the link power napping transitions. The result is a delay δ. This happens during every reduce phase of the Map reduce algorithm.

Reasons behind this Speculative diagram:

The diagram shows three portions 1. the initial loss of packets 2. incast delay and 3.transition time.

There will be an initial loss of packets due to the fact that the transition delay from the power switches is not zero.

Incast delay is attributed to handling huge number of packets suddenly Vs handling a few packets by the link during transition from near zero D2 state to the active D0 state.



CLUSTER CAST:

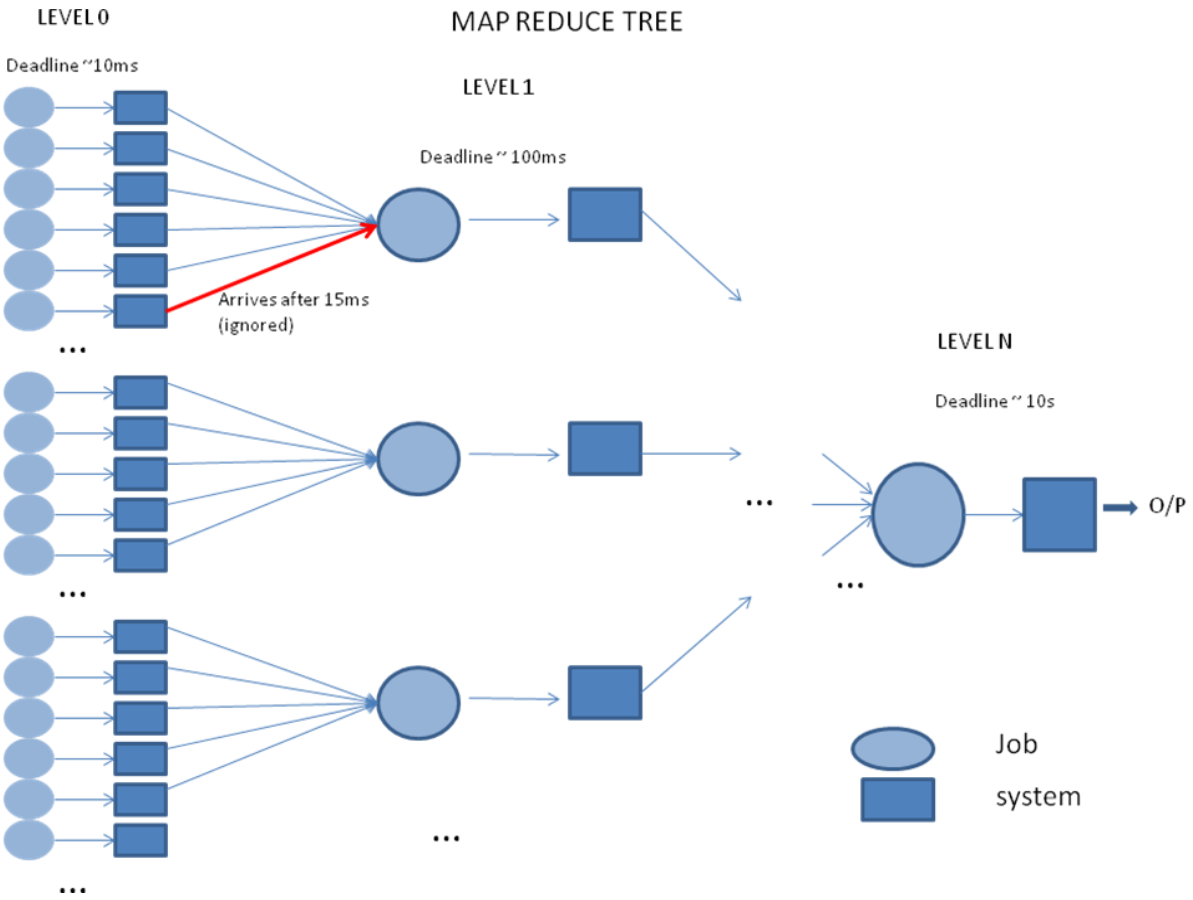
Thus the incast delay δ is introduced for each reduce stage of the iterative Map Reduce algorithm. This would results in aggregation of hardware delay across the various intermediate reduce stages in the Map Reduce Tree. This would eventually lead to a non-compliance with the map reduce deadlines at a later stage. Map reduce Algorithm ignores the jobs which doesn't meet the set deadlines. Now if the job ignored is in the lower levels of map reduce tree the diminishing quality due to that is acceptable as the jobs at the lower levels doesn't impact the quality of the result significantly. But if the intermediate Jobs of higher levels closer to the root(where the results are presented) are ignored due to non compliance of dead-line, then the degradation in the quality of result is huge. This phenomenon is called as Cluster Cast.

The following are the reasons why Power napping switches might end up in cluster Cast:

1. There is a vast number of micro delays in Data center traffic. Interactive servers are idle for about 60% idle, most of the idle intervals are under one second[11].

2.The Power Nap switches are Aggressive. They try to save power whenever possible by means of switching them off when the tasks in hand are completed as the transition time is less.

So When the Incast packets arrives in one link, it would not be surprising to find the power nap link taking nap to save power. But again the probability of Cluster cast is less due to the fact that the Synchronization of the reduce stage is necessary for AT EVERY STAGE to result in an Cluster Cast. It should be noted that a chain of incasts fully synchronized at every reduce stage results in a cluster cast. Any break in the chain would stop the formation of cluster cast from the incast. The probability of having incast reduce at every intermediate stage is low (randomness of jobs).



CLUSTER CAST: Thus if such as phenomenon occurs the incast delay (δ) of each stage gets aggregated and results in the deadline miss at a later stage in the map reduce tree and cause a huge impact in the result of the web search or a data mining operation. Due to the incast delay, the jobs at the lower levels might not miss the deadline but once such delays get aggregated over the higher levels the result is a tangible impact on the final output.

TO ALLEVIATE CLUSTER CAST:

The infiniband/YARC supported link rate adaption can be done for the Incast prone Links instead of Power Nap. Here the latency of the link rate adaption are in the order of nano seconds whereas the power nap takes 200 µs latency to wake up. thus the latency of the incast can be alleviated significantly with the advent of Infiniband link rate adaption. There is a tradeoff between the amount of energy saved between the power nap (near zero idle power 400mW with latency 200 µs) Vs YARC Link Rate Adaption (whose latency is in the order nano seconds with the power during the idle state of 80W). This would involve marking the links in the switches/routers to be incast prone. Although that can be achieved by means of profiling the links when there is incast congestion taking place. Again, this would reduce the amount of incast delay experience in the switch. But the cluster cast will occur due to the fact that there is still aggregation between the incasting switches. The number of levels to reach the state of missing the deadline by the jobs increase significantly. At the same time the probability of such an occurrence decreases significantly. And also the impact of such a phenomenon will be more devastating than the actual cluster cast. The reason is simple. it makes the job losses the deadline at a level closer to the root and hence the loss of throughput will be phenomenal!

CONCLUSION:

This project summarized the energy proportional methods , and the benefits procured administering them on the enterprise and data center networks. The extent to which energy proportionality can be achieved in the enterprise networks and the data center networks are studied, analyzed ,summarized. This project speculates the possible side effects for conventional energy proportional methods. It introduces Buffer Burst phenomenon due to energy proportional measures independently acted based on individual link statistics. It summarizes the data center traffic characteristics and the possible energy proportional methods that could work on those methods. It combines the techniques of power nap and the Link rate adaption from infiniband YARC routers to produce a highly flexible power switch which has the off power latency of 200 µs and 400mW with link rate adaption during idle states(especially for incast prone links) with the idle power of 80W and latency of 10 µs at max. It suggests the technique of Power switches for almost any Data center to get the maximum energy savings. It introduces the possible phenomenon of Cluster Cast which could happen when the Power nap switches are administered in Map reduce Data centers and tries of explain how infiniband link rate adaption could be followed to reduce the impact of the cluster cast. In spite of this, cluster cast might persist.

Acknowledgements: The author would like to thank Prof Badri Nath, Rutgers, Rekha Bachwani, PHD Student, DARK LABS Rutgers and Cheng li, PHD Student, DARK LAB, Rutgers for their valuable suggestions to improve the work.

REFERENCES:

1. Luiz André Barroso , Urs Hölzle, **The Case for Energy-Proportional Computing**, Computer, v.40 n.12, p.33-37, December 2007 [doi>10.1109/MC.2007.443]

2. Priya Mahadevan , Sujata Banerjee , Puneet Sharma, **Energy proportionality of an enterprise network**, Proceedings of the first ACM SIGCOMM workshop on Green networking, August 30-30, 2010, New Delhi, India [doi>10.1145/1851290.1851302]

3. Dennis Abts, Micheal R.Marty, Philip M.Wells, Peter Klauster, Hong Liu, **Energy proportional Data centers**, Proceedings of the 37th annual international symposium on computer architecture, ACM New York, NY, USA 2010, ISBN:978-1-4503-0053-7.

4. Theophilus Benson, Aditya Akella, David A.Maltz, **Network Traffic Characteristics of Data Centers in the Wild** , IMC '10 Proceedings of the 10th annual conference on Internet Measurement. ISBN -978-1-4503-0483-2.

5. Brandon Heller , Srini Seetharaman , Priya Mahadevan , Yiannis Yiakoumis , Puneet Sharma , Sujata Banerjee , Nick McKeown, **ElasticTree: saving energy in data center networks**, Proceedings of the 7th USENIX conference on Networked systems design and implementation, p.17-17, April 28-30, 2010, San Jose, California

6. Willis Lang, Jignesh M.Patel, **Energy Management from Map-Reduce Clusters**, Proceedings of the VLDB Endowment, Volume 3 Issues 1-2 September 2010.

7. Albert Greenberg , James R. Hamilton , Navendu Jain , Srikanth Kandula , Changhoon Kim , Parantap Lahiri , David A. Maltz , Parveen Patel , Sudipta Sengupta, **VL2: a scalable and flexible data center network**, Proceedings of the ACM SIGCOMM 2009 conference on Data communication, August 16-21, 2009, Barcelona, Spain [doi>10.1145/1592568.1592576]

8. Theophilus Benson , Ashok Anand , Aditya Akella , Ming Zhang, **Understanding data center traffic characteristics**, Proceedings of the 1st ACM workshop on Research on enterprise networking, August 21-21, 2009, Barcelona, Spain

9. Haitao Wu, Zhenqian Feng, Chuanxiong Guo, Yongguang Zhang, **ICTCP: Incast Congestion Control for TCP in Data Center Networks**, CO-NEXT '10 Proceedings of the 6th International Conference, ISBN-978-1-4503-0448-1.

10. Mohammad Alizadeh , Albert Greenberg , David A. Maltz , Jitendra Padhye , Parveen Patel , Balaji Prabhakar , Sudipta Sengupta , Murari Sridharan, **Data center TCP (DCTCP)**, Proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM, August 30-September 03, 2010, New Delhi, India.

11. B. Nordman. **Networks, energy and energy efficiency**. Cisco Green Research Symposium, March 2008.

12. William Maruyama, Mark George, Eileen Hernandez, Keith LoPresto, YeaUang, "**Enterprise Network Control and Management: Traffic Flow Models**", http://ieeexplore.ieee.org/ielx5/6639/17707/00821384.pdf?arnumber=8213844

13. A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, **Cutting the Electric Bill for Internet-Scale Systems,** ACM SIGCOMM, 2009.

14 David Meisner , Brian T. Gold , Thomas F. Wenisch, **PowerNap: eliminating server idle power**, Proceeding of the 14th international conference on Architectural support for programming languages and operating systems, March 07-11, 2009, Washington, DC, USA [doi>10.1145/1508244.1508269]

15. S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken. **The Nature of Data Center Traffic: Measurements and Analysis.** In IMC, 2009

16. F. A. Tobagi, "**Fast Packet Switch Architectures for Broadband Integrated Services Digital Networks**," Proc. IEEE, vol. 78, no. 1, Jan. 1990, pp. 133-67

17. M. Arpaci and J. Copeland, "**Buffer Management for Shared-Memory ATM Switches**," *IEEE Comm. Surveys and Tutorials*; http://www.comsoc.org/livepubs/surveys/public/1q00issue/copeland.html.

18. Wei Kuang Lai and Mu-Rung Shiu," **Improving goodputs of IP packets under ATM UBR traffic with port-based queueing schemes",** http://www.sciencedirect.com/science/article/B6VRG-40WDYC4-6/2/12170af43e340965dee6aba876d8ed23.

19. A. Jardosh et al. **Towards an Energy-Star WLAN Infrastructure**. In HOTMOBILE. 2007

20. V. Vasudevan, A. Phanishayee, H. Shah, E. Krevat, D. Andersen, G. Ganger, G. Gibson, and B. Mueller. **Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication**. In Proc SIGCOMM, 2009.

21. T. Benson, A. Anand, A. Akella, and M. Zhang. **The case for fine-grained traffic engineering in data centers**. In Proceedings of INM/WREN '10, San Jose, CA, USA, April 2010.

22. PCI 1.1 Power Specification http://www.pcisig.com/specifications/conventional/pcipm1.2.pdf

23. SMSC, "**LAN9420/LAN9420i single-chip ethernet controller with HP Auto-MDIX support and PCI interface**," 2008. http://www.smsc.com/media/Downloads/Product_Brochures/lan9420fs.pdf

24. H. Wu, Z. Feng, C. Guo, and Y. Zhang. **ICTCP: Incast Congestion Control for TCP in Data Center Networks**. In ACM CoNEXT, 2010.

25. Steve Scott, Dennis Abts, John Kim, and William J. Dally. **The blackwidow high-radix clos network**. In ISCA ’06. Proceedings of the 33rd annual International Symposium on Computer Architecture, pages 16–28, 2006.