

Flow-Through-Queue based Power Management for Gigabit Ethernet Controller

Hwisung Jung, Andy Hwang^{*}, and Massoud Pedram

University of Southern California, Broadcom Corp.*

Agenda

- Introduction
- FTQ-based Architecture
- Modeling the FTQ-based System
- SMDP-based Energy Optimization
- Multiple V_{dd}/V_{th} Assignment Algorithm
- Experimental Results
- Conclusion

Introduction

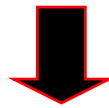
- Implications of high-functionality and high-performance design:
 - higher power densities
 - higher temperature
 - lower circuit reliability
- Gigabit Ethernet controller
 - Power increases rapidly with increase in the link speed
- Current design technologies allow:
 - Dynamic voltage frequency scaling (DVFS)
 - Multiple Vdd/Vth assignments
- Synchronization solution:
 - Globally asynchronous locally synchronous (GALS) architecture

Selected Prior Work

- A. Iyer, et al. (ICCAD 2002)
 - Voltage scaling in multiple voltage cores
- D. Lackey, et al. (ICCAD 2002)
 - Voltage islands with multi-threshold CMOS
- A. Srivastava, et al. (DAC 2004)
 - Simultaneous dual- V_{dd} and dual- V_{th} assignment
- S. Bhunia, et al. (TComp 2005)
 - Adaptive task voltage scaling for GALS
- Q. Wu, et al. (HPCA 2005)
 - DVFS scheme in multiple clock domains

Motivation

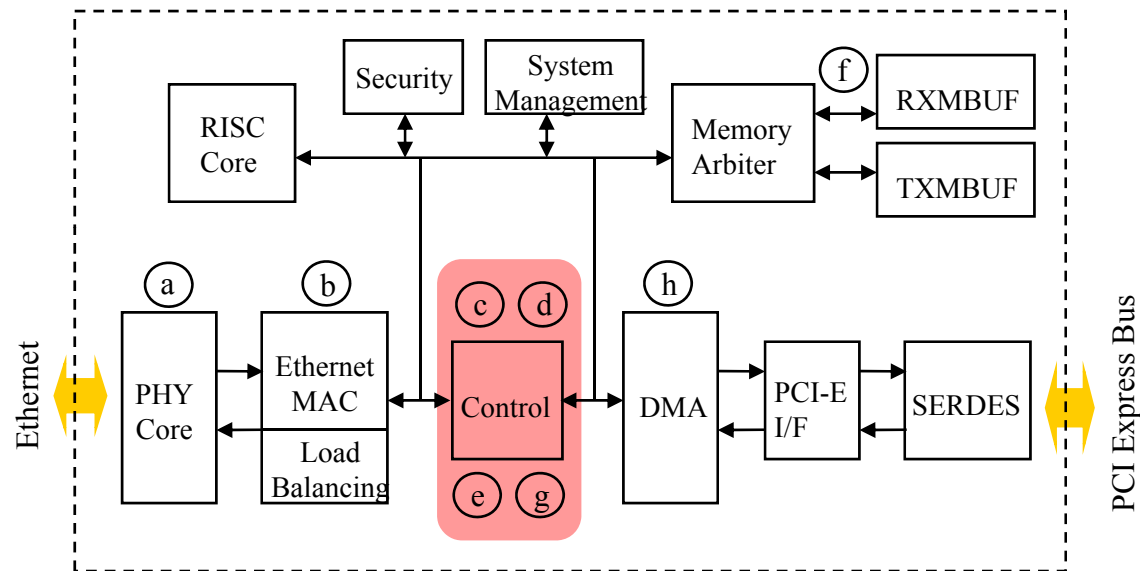
- No prior work on system-level stochastic power management w/ static V_{th} assignment and dynamic V_{dd} selection.
- GALS results in performance penalty due to complexity of the configuration.



- Systematic approach for a stochastic power management framework for V_{dd}/V_{th} assignments
- Power management architecture based on a Flow-Through-Queue (FTQ)-assisted synchronization mechanism.

Background

■ Block diagram of a Gigabit Ethernet controller

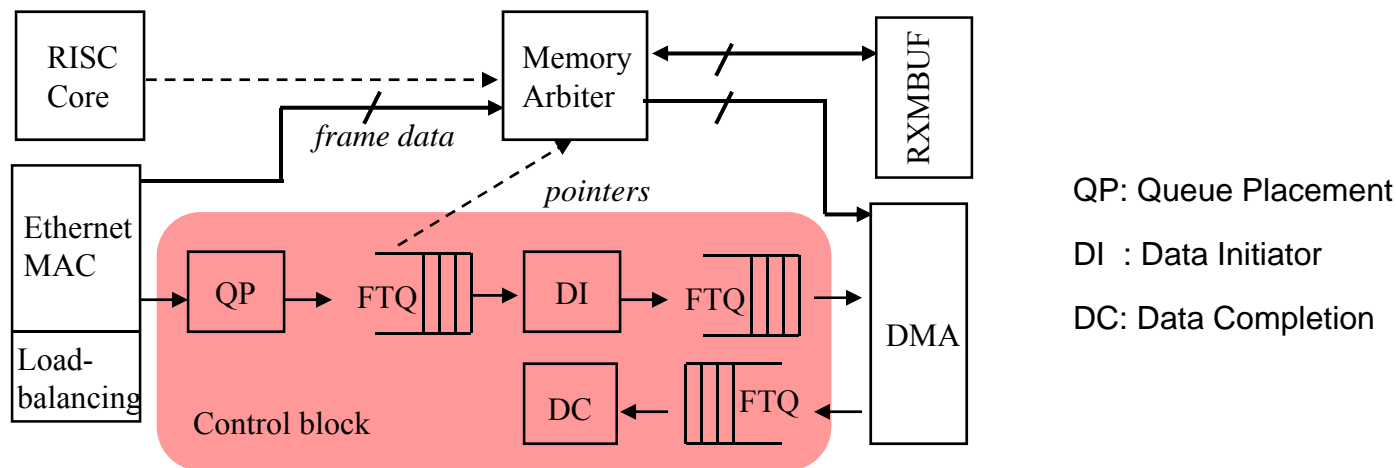


- (a) Receive a data stream from Ethernet
- (b) Perform address checking and CRC calculation
- (c) Calculate checksum and parse TCP/IP headers
- (d) Classify the frame based on a set of matching rules
- (e) Strip Virtual Local Area Network (VLAN) tag
- (f) Place packet data and header into buffer
- (g) Complete buffer descriptions for packets
- (h) DMA transfers data to the host memory

(Refer to: <http://www.broadcom.com> NetXtreme Gigabit Ethernet Controller document)

FTQ-based Architecture (1)

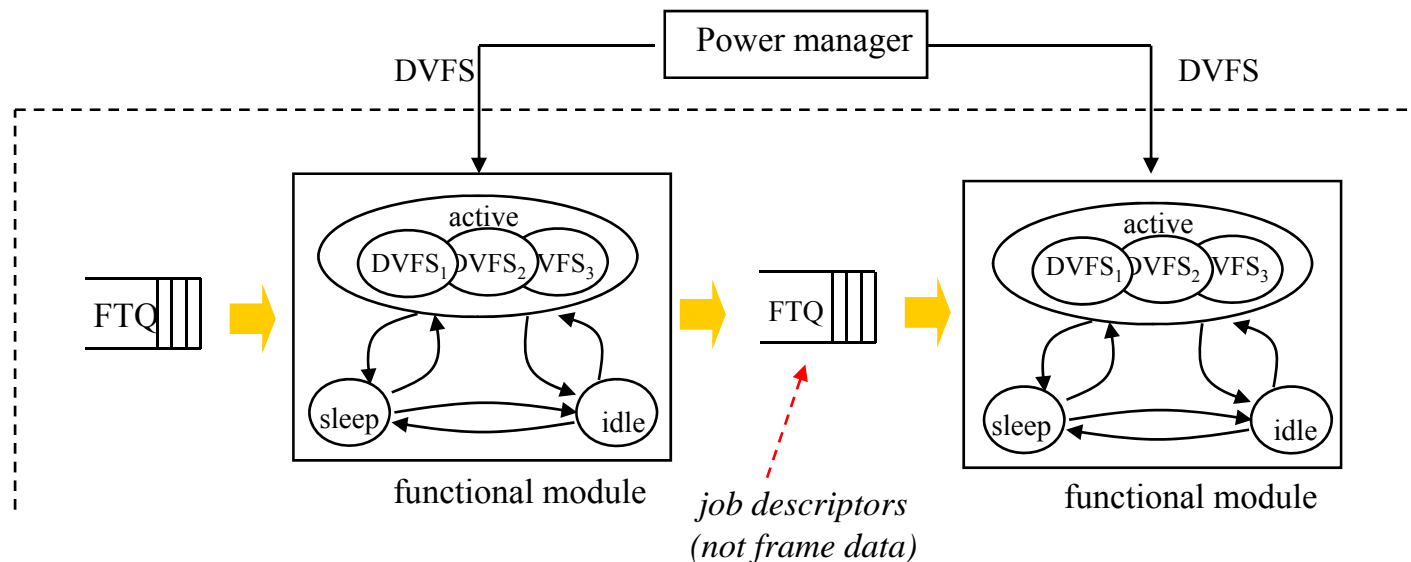
- The Flow-Through-Queue mechanism enables multiple clock and voltage levels inside the Ethernet controller
 - FTQ-based architecture provides FIFO mechanism for data transfer.
 - It deals with the control dominated tasks (c, d, e, and g), which must have low-latency. Target blocks are QP, DI, and DC.
 - State machine of each control block reacts to contents of its FTQ.



Configuration with the FTQ in the packet receive path

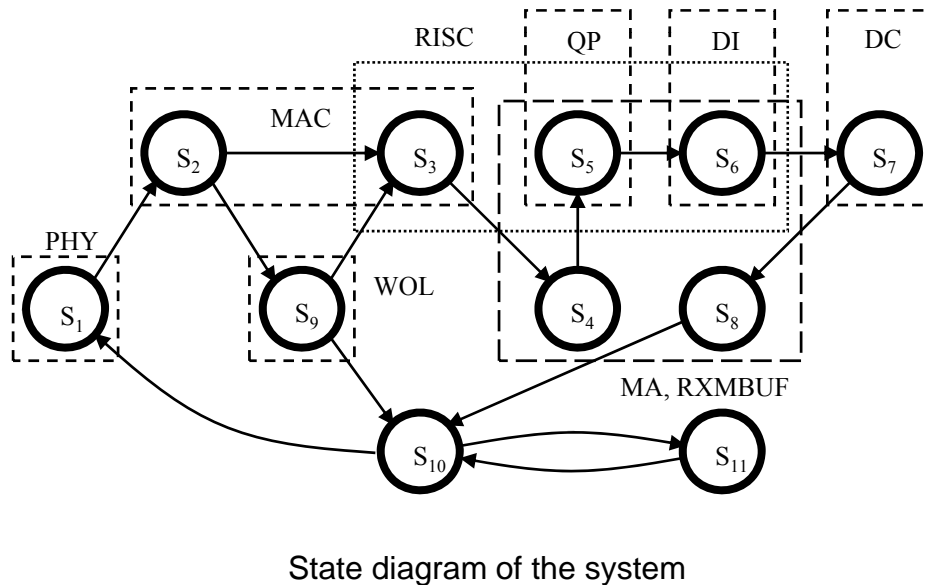
FTQ-based Architecture (2)

- Functional modules (i.e., QP, DI, and DC) can switch between different power-speed levels
 - The power manager can use information about the FTQ of each module to select the appropriate voltage and frequency setting.
 - Each FTQ contains job descriptors, which are used to indicate where the frame data is located in the buffer.



Modeling FTQ-based System (1)

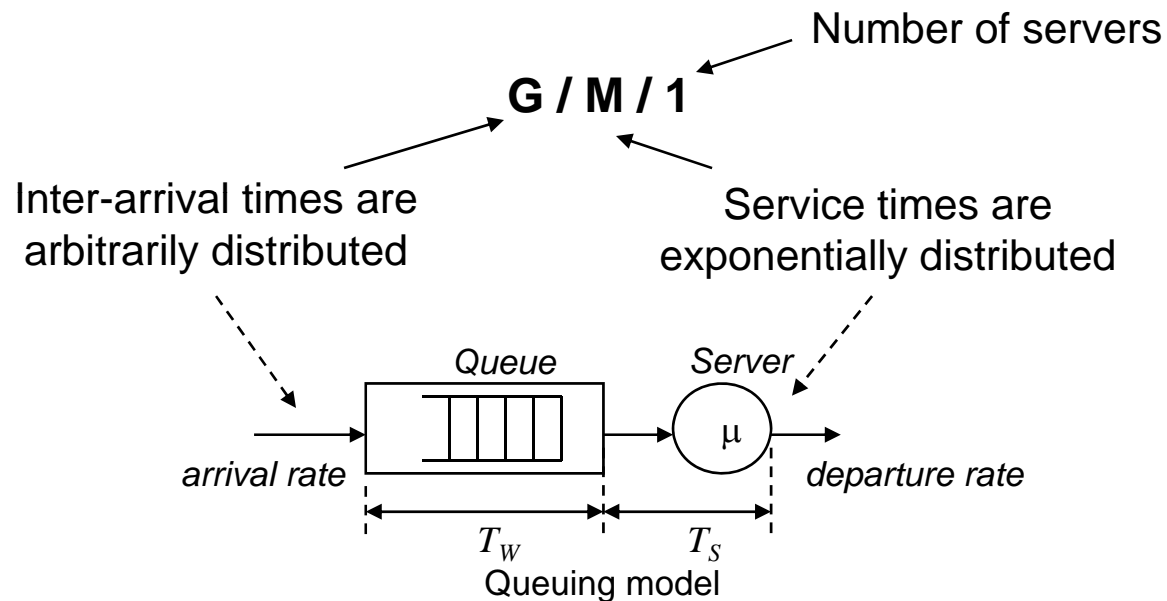
- Realistic modeling of a system is an important step toward optimizing the performance and energy consumption.
- Semi-Markov Decision Process (SMDP) model enables the user to apply mathematical optimization techniques to derive DPM policies.



State	Description
S ₁	Receive data stream from physical layer interface
S ₂	Perform address checking, CRC calculation, and CSMA/CD
S ₃	Calculate checksum and parse TCP/IP header
S ₄	Place packet data and header into buffer memory
S ₅	Buffer descriptor processing (Queue replacement)
S ₆	Buffer descriptor processing (Data Initiator)
S ₇	Complete buffer descriptor for packet
S ₈	DMA transfers packet data to host memory
S ₉	Filter WOL (Wakeup on LAN) packets during power down
S ₁₀	System idle mode
S ₁₁	System sleep mode

Modeling FTQ-based System (2)

- Each FTQ may be represented by a G/M/1 queuing model:



- The more commonly used M/M/1 queuing model underestimates the occurrence probability of requests with long inter-arrival times.

Modeling FTQ-based System (3)

- Let W denote the number of waiting tasks in the FTQ just before a new task arrives, then we have

$$q_n = Prob\{W = n\} = (1 - \gamma)\gamma^n, \quad n = 0, 1, \dots, \infty$$

where γ is the unique solution (real, $0 < \gamma < 1$) of Laplace-Stieltjes transform (LST) of the inter-arrival time distribution function.

- Let $T_{W,k}$ represent the *waiting time* in the k^{th} FTQ, then the waiting time is given by

$$T_{W,k} = \frac{\gamma}{\mu(1 - \gamma)}$$

- The *utilization ratio* of a functional module is defined as:

$$u_k = \frac{BP_k}{BP_k + IP_k}$$

where BP is duration of the busy period of the module whereas IP is its idle period.

SMDP-based Energy Optimization (1)

- Let $actpow_{k,Vdd,Vth}$ and $slpow_{k,Vdd,Vth}$ represent the power consumption in the k^{th} functional module during its active and sleep modes.
- The expected cost rate (i.e., active power dissipation) is the summation of state-dependent power term and a transition dependent energy cost:

$$cost(s, a) = \sum_{k \in K} actpow_{k,Vdd,Vth} + \frac{1}{\tau(s, a)} \sum_{s' \in S} Prob(s' | s, a) ene(s, s')$$

- K denotes the set of functional modules
- $ene(s, s')$ is the energy required by the system to transit from state s to s'
- $\tau(s, a)$ is the expected duration of the time that the system spent in the state s if action a is chosen.

SMDP-based Energy Optimization (2)

- Let a sequence of states s^0, s^1, \dots, s^k denote a processing path δ from s^0 to s^k with the property that $p(s^0, s^1), \dots, p(s^{k-1}, s^k) > 0$, where $p(x, y)$ is the probability that the system moves from state x to state y .
- For a given policy π , the average active power dissipation can be given over the set of processing paths:

$$actpow_{avg}^{\pi}(\delta) = EXP\left[\sum_{i=0}^k \phi^i cost(s^i, a^i)\right] \quad (\phi: \text{discount factor, } 0 < \phi < 1)$$

- The average energy dissipation of the module can be calculated as:

$$ene_{avg} = actpow_{avg}^{\pi}(\delta) \cdot \sum_{l \in L} \sum_{k \in K} Texe_{l,k.Vdd.Vth} + \sum_{k \in K} slpow_{k.Vdd.Vth} \cdot (T_d - \sum_{l \in L} Texe_{l,k.Vdd.Vth})$$

- L denotes the set of tasks
- T_d is the user-specified total computation time
- $Texe_{l,k.Vdd.Vth}$ is the execution time of task l on functional unit k running at V_{dd} and V_{th} .

SMDP-based Energy Optimization (3)

- The goal is to minimize energy consumption of a SMDP system, G , subject to performance constraints:

$$\min \text{ene}_{avg}$$

$$\text{s.t. } \sum_{k \in \delta} (T_{W,k} + T_{S,k}) \leq T_d \quad \forall \delta \in \text{paths}(G)$$

$$BP_k / (BP_k + IP_k) \geq u_k^* \quad \forall k \in K$$

$$T_{W,k} = \sum_{i=1}^n i \cdot q_{i,k}, \quad T_{S,k} = 1/\mu_k$$

$$BP_k = \sum_{i=1}^n q_{i,k}, \quad IP_k = q_{0,k}$$

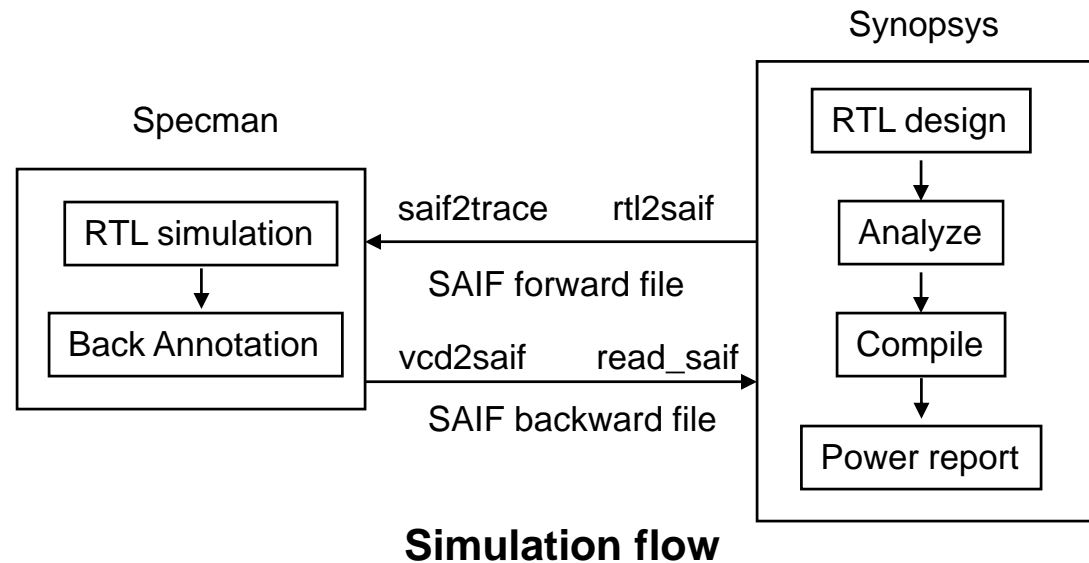
$$\sum_{i=0}^n q_{i,k} = 1 \quad \forall k \in K$$

$$0 \leq q_{i,k} \leq 1 \quad i = 0, \dots, n$$

- The service time on module k , $T_{S,k}$, is influenced by the DVFS setting
- u_k^* is a lower bound on the utilization of functional module

Workload-Aware Vdd/Vth Assignment (1)

- The multiple Vdd/Vth assignment method begins with optimizing a circuit for a maximum speed by using the available slack.
- Use TSMC130nm LP library: (1.35V, 1.5V, and 1.65V) V_{dd} and dual (High and Low) V_{th} .
- Use SAIF (Switching Activity Interchange Format) for power calculation.



Workload-Aware V_{dd}/V_{th} Assignment (2)

- A simple V_{dd}/V_{th} assignment algorithm

Determine the timing critical paths of the circuits.



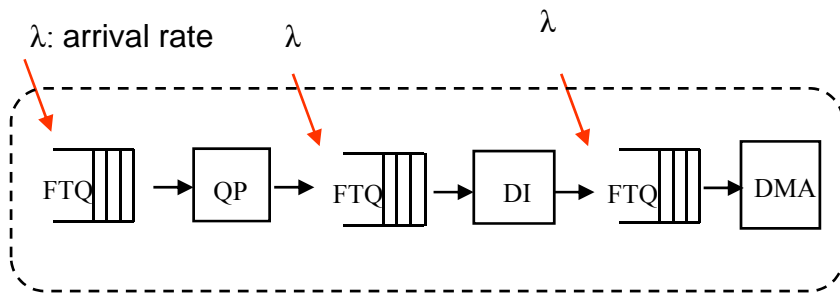
Apply high supply voltage, V_{dd,h}, and low threshold voltage, V_{th,l} for the gates of those paths.



Use low supply voltage, V_{dd,l}, for other gates which drive large capacitance.

Experimental Results (1.1)

- SMDP-based Energy Optimization
- Set the performance constraints on T_d and u_k
 - E.g., $T_d = 5$ and $u_k = 0.6$
 - Consider different task arrival rates.



Arrival rate	Vth	QP		DI		DMA		
		Vdd	Ene. Acti.	Ene. idle	Ene. Acti.	Ene. idle	Ene. Acti.	Ene. idle
$\lambda = 0.8$	Vth.h	1.35	28.5	8.1E-4	76.3	18E-4	118.1	36E-4
		1.50	35.6	2.7E-4	94.3	6.3E-4	145.8	8.1E-4
		1.65	42.5	1.8E-4	111.4	10E-4	176.2	7.9E-4
	Vth.l	1.35	28.4	17E-4	76.2	35E-4	118.3	13E-3
		1.50	35.6	4.5E-4	94.5	10E-4	144.9	38E-4
		1.65	42.6	6.3E-4	111.2	16E-4	176.1	55E-4
$\lambda = 0.7$	Vth.h	1.35	18.2	9.0E-4	48.7	20E-4	75.4	41E-4
		1.50	22.8	3.2E-4	60.0	7.2E-4	93.1	8.0E-4
		1.65	27.1	2.1E-4	72.1	11E-4	112.4	9.2E-4
	Vth.l	1.35	18.2	19E-4	48.7	39E-4	75.5	15E-3
		1.50	22.7	4.9E-4	60.1	13E-4	93.2	43E-4
		1.65	27.0	7.1E-4	72.3	18E-4	112.4	61E-4
$\lambda = 0.6$	Vth.h	1.35	13.4	1.1E-4	36.0	22E-4	55.7	44E-4
		1.50	16.8	3.0E-4	44.5	8.1E-4	68.9	9.1E-4
		1.65	20.1	2.2E-4	54.1	13E-4	83.1	10E-3
	Vth.l	1.35	13.4	21E-4	36.2	42E-4	55.7	15E-3
		1.50	16.7	6.0E-4	44.5	13E-4	68.8	57E-4
		1.65	20.1	8.3E-4	54.1	18E-4	82.9	70E-4

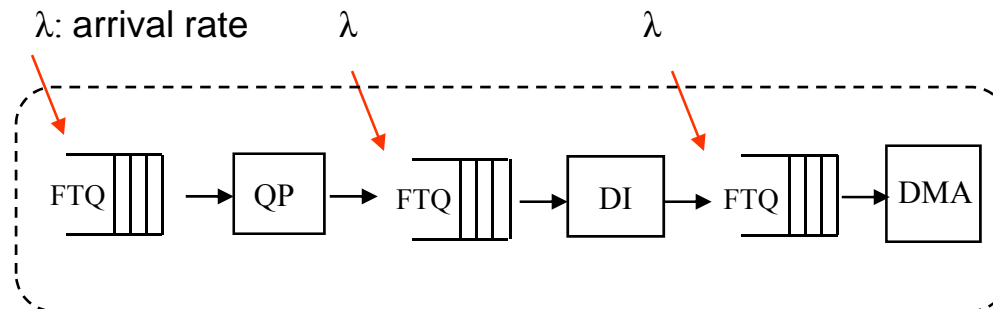
Energy dissipation for various workloads (normalized)

Experimental Results (1.2)

- Consider combinations of different workloads for each module
 - Achieve energy savings for both active and idle modes up to 20% and 56%, respectively.

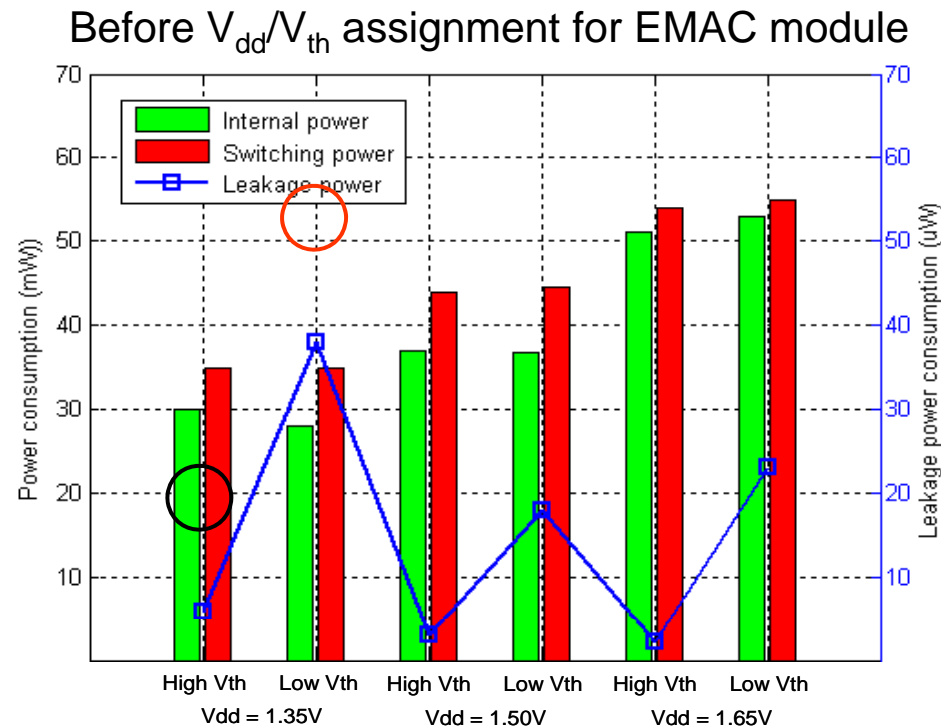
Workload: arrival rate (λ)			Total energy (typical)		Proposed policy		Savings	
QP	DI	DMA	active	idle	active	idle	active	idle
0.8	0.7	0.6	164.5	53E-4	132.4	24E-4	20%	55%
0.7	0.6	0.5	123.9	78E-4	100.1	34E-4	20%	56%
0.6	0.7	0.8	222.6	77E-4	180.0	36E-4	19%	53%
0.5	0.6	0.7	151.4	63E-4	122.4	39E-4	19%	54%

Energy optimization for various workloads (normalized)



Experimental Results (2.1)

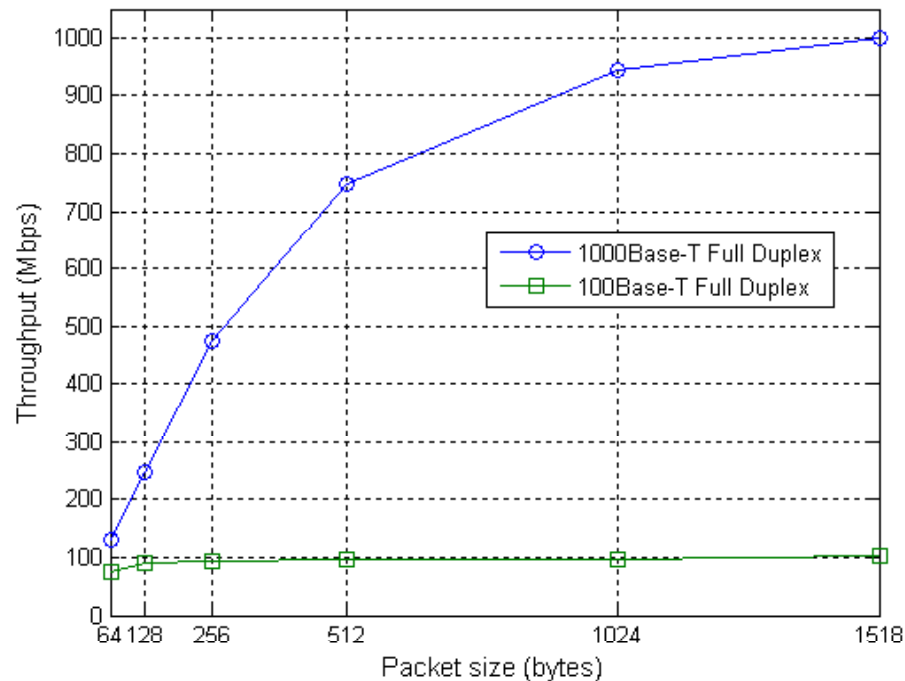
- Workload-Aware V_{dd}/V_{th} Assignment



- All-V_{th,h} cell-based design consumes 5.8uW of power with 16.2ns latency.
- All-V_{th,l} cell-based design consumes 38uW of power with 9.36ns latency.

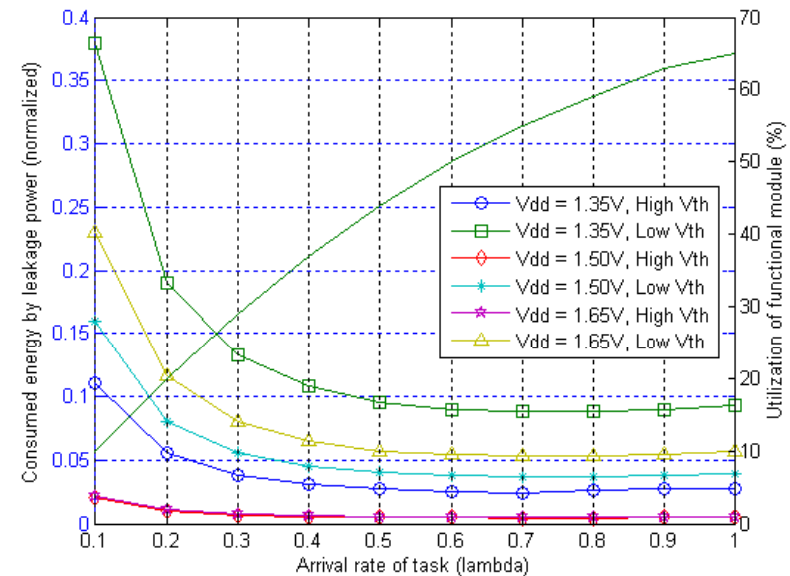
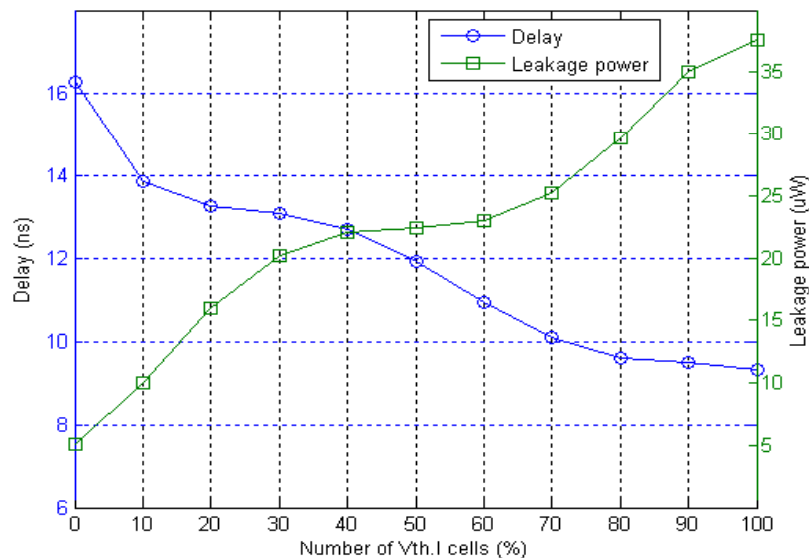
Experimental Results (2.2)

- Performance characteristics
 - Maximum 100Base-T and 1000Base-T full duplex bandwidths for each packet size are achieved.
 - The IP packet size is varied; The inter-packet gap is kept at 0.0096us.



Experimental Results (2.3)

- We focus on energy consumption due to leakage currents in the idle mode of module and on the total computation time
 - Calculate the utilization ratio of the target module (e.g., EMAC)
 - This method can adjust the V_{dd} value when the workload characteristics change.



Conclusion

- With knowledge of the applications and their requirements, DPM provides the flexibility to reduce voltage and frequency to minimal levels.
- Fine-grained power management method results in significant energy savings for various workload under performance constraints.
- Performance optimization problem based on the SMDP and DVFS were formulated and solved.
- Simulation results demonstrate system-wide energy savings for both active and idle modes up to 20% and 56%, respectively.