Adaptive Power Management for the On-Chip Communication Network

Guang Liang
University of Amsterdam, The Netherlands

Axel Jantsch
Royal Institute of Technology, Sweden

***Abstract*— An on-chip communication network is most power efficient when it operates just below the saturation point. For any given traffic load the network can be operated in this region by adjusting frequency and voltage. For a deflective routing network we propose the design of a central controller for dynamic frequency and voltage scaling. Given history information including the load and frequency in the network, the controller adjusts the frequency and voltage such that the network operates just below the saturation point. We provide control mechanisms for continuous and discrete frequency ranges. With a discrete frequency range and taking into account voltage switching delays, we evaluate the control mechanism under stochastic, smoothly varying and very bursty traffic. Experiments demonstrate that adaptive control is very effective in minimizing power consumption at reasonable performance. Compared with a fixed high frequency network, the adaptively controlled network is significantly more power efficient. We compare it to fixed frequency networks, which are either too slow exhibiting unbounded delays, or are dimensioned for the worst case with very high frequency and are very power hungry.**

I. Introduction

Dynamic Voltage Scaling (DVS) has been successfully applied to microprocessors [1] to minimize power consumption while achieving acceptable performance. When the processor runs below its maximum working load, the clock frequency can be decreased until there are no idle cycles and the processor is fully loaded measured in useful operations per clock cycle. A lower frequency allows for a lower operating voltage. Decreasing both frequency and voltage can significantly reduce the power consumption since power consumption is proportional to the clock frequency and to the square of the voltage. Thus, by operating the processor as slow as possible while still meeting all timing constraints the power consumption can be minimized.

In networks on chip the same principle can be applied to the communication network. It will consume the least power when operated as slow as possible while still meeting all timing constraints. Interconnects are a major, if not the dominant source of energy consumption in today's systems-on-chip [2]. In networks with simple switching schemes, power consumption is dominated by the link power on the channels between switches. Shang et al. [3] report that 82.4% of the network power is consumed by the links in a 2-dimensional mesh topology with wormhole routing, 32 bit flits, 2 virtual channels and 128 flit buffers per input port. Vitkowski et al. [4] report a 98% share of the network power consumption by the links in a deflective routing network with switches that have no internal packet buffers. The communication bandwidth in future network on chip architectures will probably be only limited by prohibitive levels of power consumption [5].

We apply dynamic voltage and frequency scaling to the entire network of the Nostrum NoC. Nostrum is a regular 2-D mesh network with a deflective routing scheme, minimal switch internal buffering and very high link level bandwidth with 128-bit buses [6], [7], [8]. It assumes a fairly regular layout that allows to predict and control the switch-to-switch signal delays. All switches have a common clock signal. Hence, they experience an identical clock frequency but different phases. The phase difference between neighboring switches can be predicted and controlled. These assumptions allow a quasi synchronous operation of the entire communication network [9]. Hence, the frequency of individual links cannot be adjusted independently. Consequently, we propose a central dynamic frequency and voltage scaling scheme that is applied to the entire network simultaneously.

The rest of the paper is organized as follows: Section II lists the most relevant works previously done on this topic; Section III describes the assumptions, parameters and formulas involved in follow-up experiments. In Section IV, we demonstrate our theory that the network runs most efficiently just below the saturation point. With this theory, in Section V, we explain the construction of the adaptive controller for both continuous and discrete frequency range. In Section VI, we show and analyze the experiment results on Nostrum with discrete frequency and voltage supply. We conclude our findings in Section VII.

II. Related Work

A variety of techniques can minimize power consumption of on-chip communication. Raghunathan et al. [2] provided an overview of techniques operating at the circuit level, the architecture level, the network level and the system level. DVS, which has mostly been used with processors and other computation units, has recently also been applied to tune power consumption and performance of interconnects.

Wei et al. [10] and Kim et al. [11] proposed DVS for off-chip communication links demonstrating a 10x power reduction potential.

Shang, Peh and Jha [3] applied the same technique to links in on-chip communication networks. They present a technique to adjust the voltage and frequency of individual links for load observed on the link during a history window.

In their 2-dimensional mesh network each router is connected over asynchronous links to its two to four neighbors. Since each link is asynchronous, its frequency can be adjusted independently of other network activities. Shang et al. reported a 3.2x average power saving with a a moderate 27.4% latency increase and 2.5% throughput reduction.

Kim et al. [12] investigated Dynamic Link Shutdown techniques, Soteriou and Peh [13] proposed on/off links together with routing and a methodology, and Worm et al. [14] proposed self-calibrating links.

All this work optimizes individual links while we focus on the entire, quasi-synchronous network. The advantage of local, link level optimization is that different links can respond more flexible to locally different demands. The disadvantages are that the overhead is expected to be much higher since control circuit is needed for every link, the network cannot effectively respond to global traffic demand, and local adaptation may lead to oscillation problems in the network with potentially unstable conditions.

We derived the estimation of the switching power consumption from [15] and switching delay from [16].

III. Assumptions and Experimental Setup

A. Nostrum Network Structure

The Nostrum NoC targets 65 nm technology and beyond. It is a regular two dimensional network, where each resource connects to exactly one switch. The switches share the same frequency with fixed phase differences. And they perform hot-potato deflection routing with minimal switch internal buffering. The network implements a best effort communication service. Nostrum also offers guaranteed latency services based on virtual circuits[8], but in our analysis here we only focus on the best effort service. The switches are connected over 128 bit buses that are approximately $2mm$ long assuming each resource is a $2mm \times 2mm$ block.

B. Frequency and Voltage Switching

Under ideal conditions, network frequency and supply voltage can be chosen on a continuous range so as to offer the best accuracy for network control. But more practically at present, the network frequency and voltage are chosen on discrete range with multiple prefixed values. The choice of frequency levels depends on the needs of particular cases.

During the switching of frequency, supply voltage has to be settled at least on the minimum value needed for the present frequency. We assume the frequency can be applied to the network immediately, but the voltage on the circuit will only change gradually, in particular when the nodes are scattered as in Nostrum. When the frequency is to be increased, supply voltage needs to be increased first. When the frequency is to be reduced, it can be decreased immediately.

In our experiment, we assume that data can be transmitted over a link while the voltage is being changed, which may be realistic as indicated by Shang et al. [3].

C. Measurement of Power Consumption

We calculate the power consumption for the links between the switches using (1).

$$\begin{aligned} P_{Link} &= \frac{1}{2}\alpha 128 C_W V_{dd}^2 f \\ C_w &= \epsilon \frac{w_{Net} L_R}{t_{ox}} \end{aligned} \quad (1)$$

where C_w is the switch capacitance for one wire, α is the switching probability of one wire, w_{Net} and L_R are the width and length of a wire, respectively, t_{ox} is the distance between wire and ground, V_{dd} is the supply voltage, and f is the frequency. The constants for 65nm technology are taken from the SIA technology roadmap, while L_R is assumed to be $2mm$.

We scaled the result for an expected 65 nm technology from the parameters of 180nm technology in UMPC180 library [17], [5]:

$$P_{65} = P_{180} \times \frac{C_{id65} V_{dd65}^2}{t_{clk65}} \times \frac{t_{clk180}}{C_{id180} V_{dd180}^2} \quad (2)$$

Voltage scaling due to frequency changes is based on [18]:

$$t_{inv} = \frac{L_d \times K_6}{(V_{dd} - V_{th})^\alpha} \quad (3)$$

where t_{inv} is the delay of one inverter, L_d, K_6 and α are technology constants and V_{th} is the threshold voltage. The switching of voltage consumes extra power [15]:

$$Energy_{overhead} = C \times (1-u) \times \left| V_{dd2}^2 - V_{dd1}^2 \right| \quad (4)$$

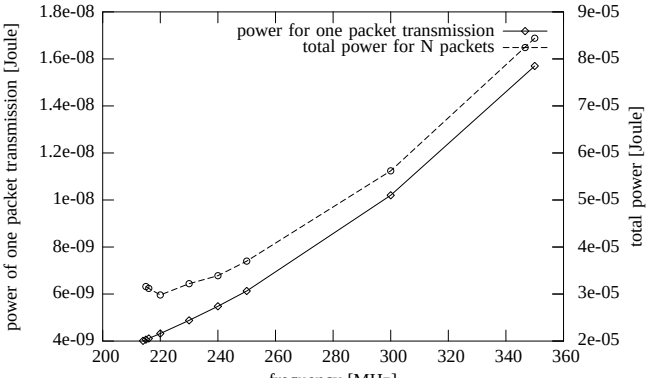
where C is the filter capacitance of the power-supply regulator on the circuit, and u is the power efficiency. In our experiments, we assume C to be $5u$F and u to be 90%. These figures are consistent with those of a single chip used in [3]. V_{dd2} and V_{dd1} are the voltages before and after the switching. Despite that the voltage on the links change gradually, we use the supply voltage to calculate power consumption.

D. Traffic Patterns for Experiments

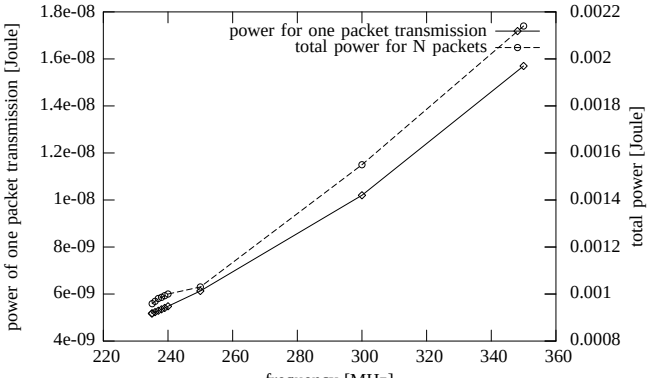
In our simulations we use three traffic types: a stochastic workload with constant emission probability, traffic with linearly changing emission probability, and traffic described by the b-model distribution [19] modeling very bursty and self-similar traffic which is typical for many communication networks. Further, we assume spatial uniform distribution of traffic.

IV. Network Power Efficiency

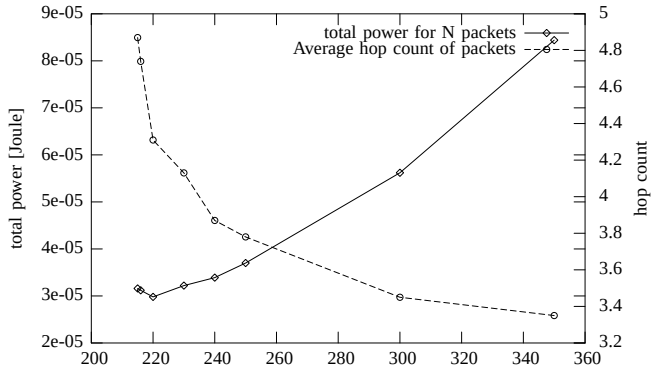
An on-chip communication network operates most power efficiently close to but below the saturation point. When the network works below the saturation point, the network can still accommodate more packets without degrading the performance. However, if the network is overloaded, extra input packets will get deflected and the packet delay rises exponentially with the load. We can adjust the load by adjusting the frequency and voltage.



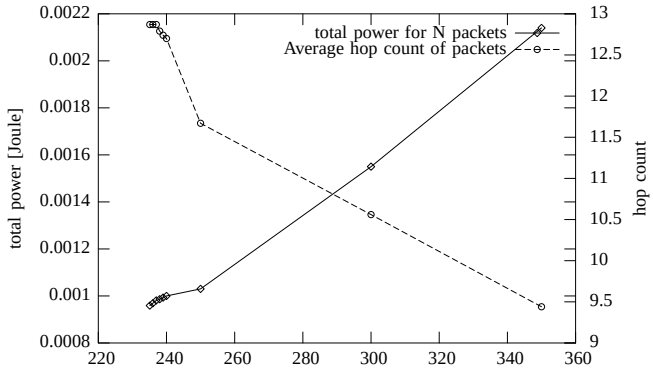
(a) a 4x4 network with 1600 packets in 1.0μ second



(b) a 12x12 network with 14400 packets in 1.94μ seconds



(c) a 4x4 network simulated with 1600 packets for 1μ second



(d) a 12x12 network simulated with 14400 packets for 1.94μ second

Fig. 1. Power consumption and hop count for a given number of packets under varying clock frequency

To substantiate our hypothesis we have simulated Nostrum networks between 4x4 and 16x16. In each simulation N packets had to be delivered within T seconds, where both N and T depend on the network size. For each given frequency we derived the minimum emission probability of each resource which would still fulfill the objective. Fig. 1 shows the result for a 4x4 and a 12x12 network.

The full line in Fig. 1(a) and 1(b) depicts the energy of transmitting one packet over a link, which decreases steadily with diminishing frequency. The total energy consumption for N packets, shown by the dotted line, also decreases until the frequency is so low that the network reaches the saturation point. When the workload is over the saturation point, more energy is consumed due to congestion. When the frequency is too low, all N packets cannot be transmitted in the given period, and the curve is cut off. For instance, the minimum frequency to deliver 1600 packets in 1μ second in a 4x4 Nostrum network is 214 MHz with an average emission probability of all resources of 95%.

Fig. 1(c) and 1(d) show again the total energy. They also depict the average hop count of packets. Although the hop count increases by 30% to 50% in the given frequency range, the total energy still decreases due to decreasing frequency. Thus, in summary it is worthwhile to operate the network as close as possible to the saturation point as the performance requirements allow.

We define the saturation point as the maximum network load when packet delay stabilizes on an acceptable figure during transmission. This point is dependent on network topology, traffic distribution and timing constraints. Our experiments suggest a proper figure as 0.5 (the ratio of network load to maximum load) for the 8x8 Nostrum network.

V. Adaptive On-chip Communication Control

We analyzed the relation between network parameters, and based on the analysis, we propose a frequency adaptation scheme which is more effective for continuous frequency change. Considering the fact that discrete frequencies are more practical in real implementation, we simplified the scheme for discrete frequency and voltage range. In general, the adaptive control is triggered by the difference between prediction of network load and the saturation point.

A. Network Parameter Analysis

Despite of the fact that there is some unpredictability in network behavior, there exist general relations between major parameters similar to the transfer functions in a typical dynamic system.

For any NoC model, (5) should always hold.

$$L_i + \sum_{j=0}^{k-1} Input_{i+j} - \sum_{j=0}^{k-1} Output_{i+j} = L_{i+k} \qquad (5)$$

L_i is the network load in cycle i. $Input$ and $Output$ are the number of packets accepted into and emitted out of the network, respectively. Equation (5) picks a window of k cycles to study the dynamics of the network. For an $N_x \times N_y$ 2-D mesh network, (5) can be modified to (6).

$$L_i + N_x \times N_y \times emissionProbability \times \frac{F_{Res}}{F_N} \times k - \sum_{j=0}^{k-1} Output_{i+j} = L_{i+k} \quad (6)$$

Without loss of generality we assume that all the resources operate at a fixed frequency F_{Res}, and F_N is the frequency of the network. In order to keep both L_i and L_{i+k} at the saturation point, based on (6), network frequency should be set to

$$F_N = N_x \times N_y \times emissionProbability \times F_{Res}/Output_i \quad (7)$$

The emission probability of resources varies and hence the traffic on the network changes correspondingly. It will be very helpful if we can estimate the emission probability so as to adjust the network control. In order to estimate the emission probability, we examine the buffer number in RNI (resource-network interface), where the packets stay waiting to be accepted into the network. We name the buffer number as RNI load. Intuitively we have

$$RNILoad_i + \sum_{j=0}^{k-1} Input_{resource2interface\_i+j} - \sum_{j=0}^{k-1} Output_{interface2network\_i+j} = RNILoad_{i+k} \quad (8)$$

When the network has constant load, the number of packets from RNI into the network should equal the number emitted out of the network. Thus, we get

$$emissionProbability = ((RNILoad_{i+k} - RNILoad_i) + \sum_{j=0}^{k-1} Output_{i+j}) \times F_N/(N_x \times N_y \times F_{Res} \times k) \quad (9)$$

if we assume equal emission probability between cycle i and $i + k$, which is practical if k is small enough. $Output$ refers to the packet number leaving the network.

Another network parameter that we need to consider is the delay introduced by sampling and computation. In more detail, there are three major sources of delay to be considered in the control mechanism, which are sampling delay, the delay of frequency computation, the delay of frequency and voltage switching. In our experiments, sampling and computation delay is predictable and quite short compared to the delay of voltage switching, so only voltage switching delay is considered explicitly.

Moreover, instead of comparing present network load to the saturation point, the controller should predict the load a few cycles after the switching, because network load changes gradually when the frequency is modified. We used linear extrapolation for network load prediction.

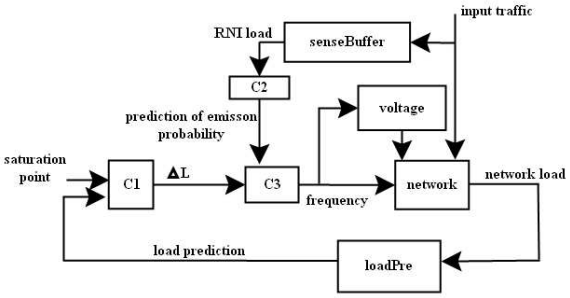


Fig. 2. Control system under continuous frequency range

B. Control Mechanism under Continuous Frequency Range

When the network on chip is supplied with continuous frequency inputs (or multiple supplies of small intervals) and voltage switching is faster than the change rate of traffic, we can build the frequency controller to return quick and very accurate response. Fig. 2 is the block diagram of the central controller. C1 is a comparator which compares predicted network load and the saturation point and returns the difference to C3. C3 combines the prediction of emission probability and the load difference, and multiplies with certain constants specific to the network to determine the frequency. The emission probability is predicted by C2 based on RNI load and network load. The frequency computed by C3 will be provided to the network as well as corresponding voltage supply. Block sensebuffer senses the buffer number in RNI and returns RNI load figure. Block loadPre predicts the network load a few cycles afterward based on linear prediction of latest network load figures.

C. Control Mechanism under Discrete Frequency Range

On Nostrum, we assume discrete frequency and voltage range, and voltage switching is much slower than potential traffic change. In particular, several levels of frequency inputs are provided to the network with corresponding voltage supplies. In this case, control mechanism for continuous frequency may not be directly applied because small frequency change will be ignored in discrete frequency range. Moreover, when we assume considerable delay in voltage switching, fast response from control system does not have actual effect. For discrete frequency range and slow switching rate, we can simplify the control mechanism as follows:

Select Case {*loadPre(network load prediction),*
satPoint(saturation Point),
epPre(prediction of emission probability)}
epPre≤0
then frequency=lower(present frequency)
loadPre-satPoint>min1
then frequency=upper(present frequency)
satPoint-loadPre>min2
then frequency=lower(present frequency)
otherwise

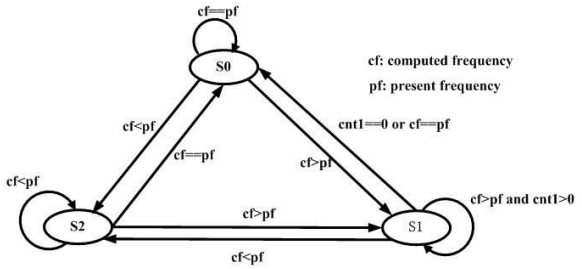


Fig. 3. State transition in frequency and voltage switching. S0 represents the state when computed frequency is equal to present frequency; S1 represents the state when computed frequency is larger than present frequency and the voltage is being lifted; S2 represents the state when computed frequency is smaller than present frequency so both the voltage and frequency are to be scaled down.

frequency= present frequency

When the prediction of emission probability is less or equal to zero, there will be no packet emission in the near future, so we lower present frequency immediately. $min1$ is the minimum difference that will increase the frequency when the load prediction is larger than saturation point; $min2$ is the minimum difference that will reduce the frequency when the load prediction is smaller. $upper$ and $lower$ are two functions which return the next frequency level above and below present frequency respectively.

The control mechanism is complicated by the inherent delay of voltage switching. As mentioned in section III-B, the voltage needs to be increased before the frequency can be lifted but there is no such need when they are decreased. We model the whole process as a state machine described in Fig. 3.

State s0 represents the state when the computed frequency is the same as present frequency, so there is no need to change the frequency. State s1 represents the state when the computed frequency is higher than present one, and the voltage is being lifted. And at the end of the state the voltage on the circuit reaches the required value for the frequency on the above level, the frequency will be lifted and the state machine goes to state s0. In both s0 and s1, if the computed frequency is smaller than present frequency, both frequency and voltage will be scaled down to the level below, and goes to state s2. State s2 represents the state when the frequency and voltage need to be scaled down. cnt1 is the number of cycles it needs to scale the voltage to the above level.

VI. Results of Experiments

In order to demonstrate the effectiveness of the adaptive control mechanism, we have done several experiments with different types (section III-D) on a Nostrum simulator. Here we provide the results for an 8x8 mesh and the following traffics: uniform traffic with 30% emission probability, uniform traffic with 70% emission probability, traffic with linearly changing emission probability (Fig. 4), and b-model traffic with b as 0.3 and average emission probability 50% (Fig. 5). Traffics with other parameters of the three patterns were also tested in the experiments but omitted in the paper.

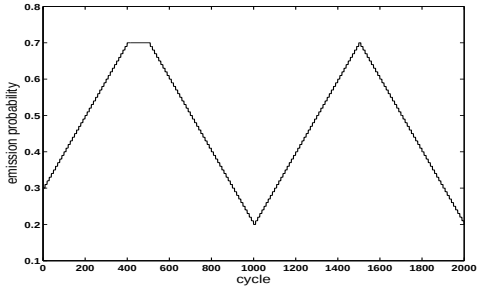


Fig. 4. Traffic with linearly changing emission probability

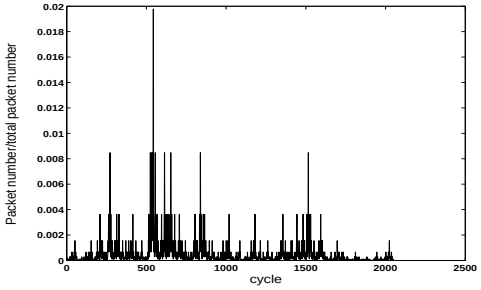


Fig. 5. b-model traffic with b=0.3 and average emission probability 50% on 2048 cycles

We used fixed frequency control and adaptive control for comparison. We chose two fixed frequencies: 60MHz and 300MHz. The adaptive control follows the mechanism described in section V-C with discrete frequency inputs: 10,30,60,90,120,150,180,210,240,270,300,350,400MHz. Saturation point (satPoint) was chosen as 0.5 as mentioned in sectionIV; future network load and emission probability (loadPre and epPre) were based on linear prediction of present figures considering a proper delay. The voltage switching speed was chosen as $2V/\mu s$[16]. The resources run at 100MHz.

We compared the network load, average energy consumed for transmitting one packet, and packet delay (from the source resource till the destination resource) for fixed frequency control and adaptive control.

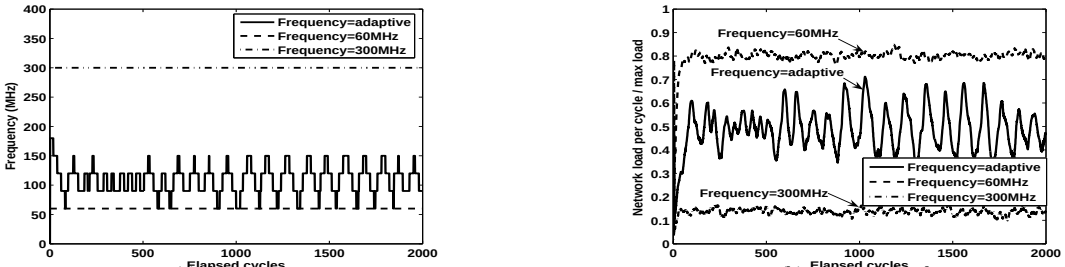
A. Frequency and Network Load

Fig. 6 to Fig. 9 depict the frequency and network load during 2000-cycle simulation of different types of traffic. The data of network load has been smoothed by averaging every 10 cycles to make the plots clearer.

We can see that adaptive frequency varies between high and low frequencies to adjust the network load around the saturation point 0.5. Frequency variation may seem redundant under uniform traffic, but as traffic pattern is subject to change from time to time, proper level of variation makes the controller respond faster to potential traffic change without costing much extra power consumption. For fixed frequencies, on the other hand, 300MHz keeps the network load too low with far greater power consumption (section VI-B) while 60MHz overloads the network with poor transmission performance (section VI-C).

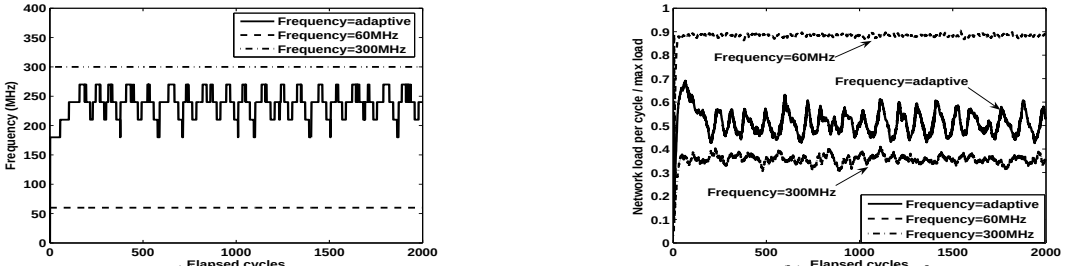
B. Power Consumption

Fig. 10 shows the average energy consumed for transmitting one packet for each traffic pattern during simulation of 2000 cycles. We observe that under adaptive control, the network



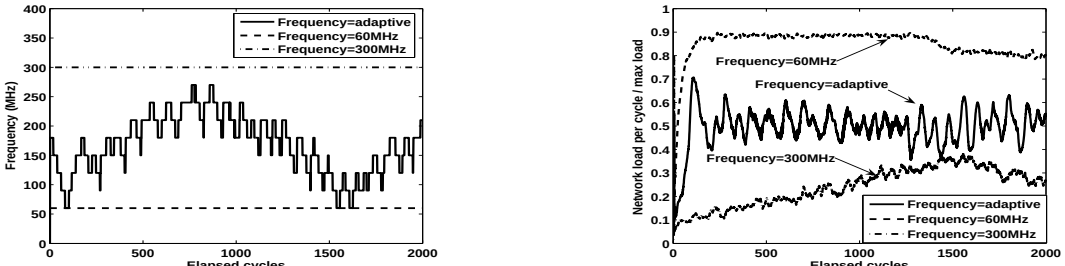
(a) frequency (b) Network load

Fig. 6. Network frequency and load under adaptive and fixed frequency for uniform traffic when emission probability is 30%



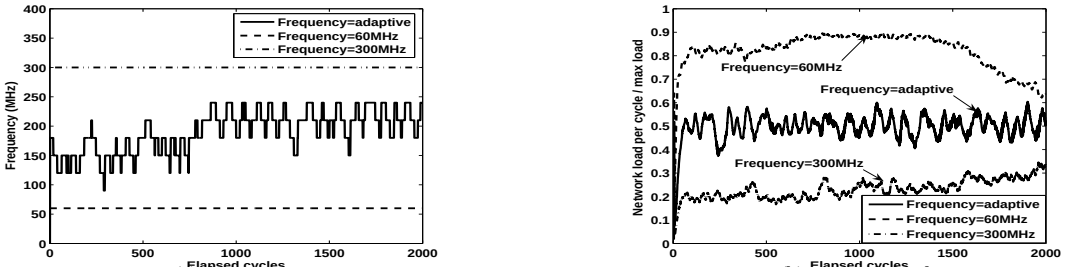
(a) frequency (b) Network load

Fig. 7. Network frequency and load under adaptive and fixed frequency for uniform traffic when emission probability is 70%



(a) frequency (b) Network load

Fig. 8. Network frequency and load under adaptive and fixed frequency for linearly changing traffic



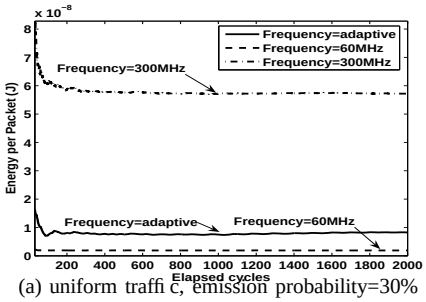
(a) frequency (b) Network load

Fig. 9. Network frequency and load under adaptive and fixed frequency for b-model traffic as in Fig. 5

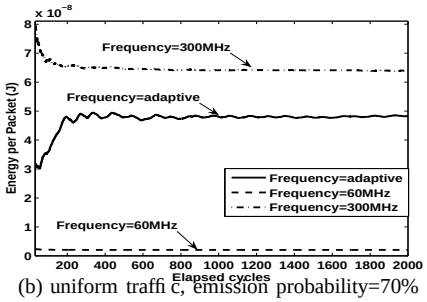
consumes much less power for transmitting the packets compared to fixed 300MHz frequency as both frequency and voltage are scaled down when possible. Although 60MHz fixed frequency consumes less power compared to adaptive control, network performance is too poor to be acceptable (section VI-C). The power consumption for adaptive control includes the extra power for voltage switching, which based on estimation and experiments, accounts for less than 10% of total consumption.

C. Packet Delay

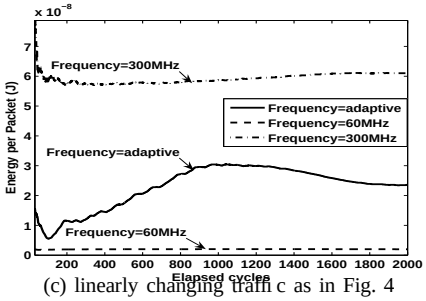
Fig. 11 illustrates the packet delay under adaptive and fixed frequency control for different traffic patterns. We examine 6000 packets for each traffic pattern. The delay is measured from the time the packet is generated by the source resource till the time it is received by the destination resource. This delay may be very different from the time the packet spends in the network, as a packet can be queuing in the interface before it is accepted to the network. In Fig. 11, x-scale is the



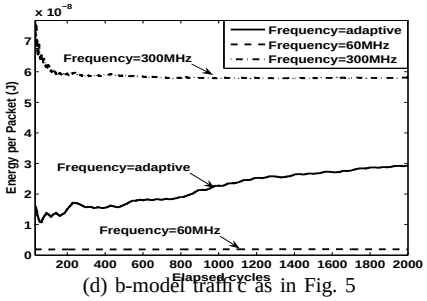
(a) uniform traffic, emission probability=30%



(b) uniform traffic, emission probability=70%



(c) linearly changing traffic as in Fig. 4



(d) b-model traffic as in Fig. 5

Fig. 10. Average packet power consumption for each traffic pattern under adaptive and fixed frequency control

TABLE I

COMPARISON OF PACKET DELAY UNDER FIXED AND ADAPTIVE FREQUENCY NORMALIZED TO THE 300MHZ CASE.

traffic	300MHz	60MHz	Adaptive
uniform,30%	1(stabilized)	16.9(accumulating)	3.17(stabilized)
uniform,70%	1(stabilized)	10.7(accumulating)	1.31(stabilized)
linearly changing	1(stabilized)	6.26(accumulating)	1.87(stabilized)
b-model	1(stabilized)	14.8(accumulating)	2.66(stabilized)

TABLE II

COMPARISON OF POWER CONSUMPTION UNDER FIXED AND ADAPTIVE FREQUENCY NORMALIZED TO THE 300MHZ CASE.

traffic	300MHz	Adaptive
uniform,30%	1	0.1441
uniform,70%	1	0.7531
linearly changing	1	0.3869
b-model	1	0.5034

indexing of packets ordered by the time they are received, and y-scale is the average delay of previous 30 received packets.

Under adaptive control, average packet delay stabilizes to a small figure as more packets are received, which is quite close to that under 300MHz fixed frequency. But under 60MHz fixed frequency, the delay accumulates rapidly with more packets are received. Considering the data is the average of 30 latest figures, the big variation of packet delay under 60MHz indicates more dramatic delay differences between packets, which may cause great trouble for the destination resources. It is attributed to the fact that when the network is overloaded, many packets in the network will get deflected and there will be a long queue in the interface as well.

D. Experiment Summary

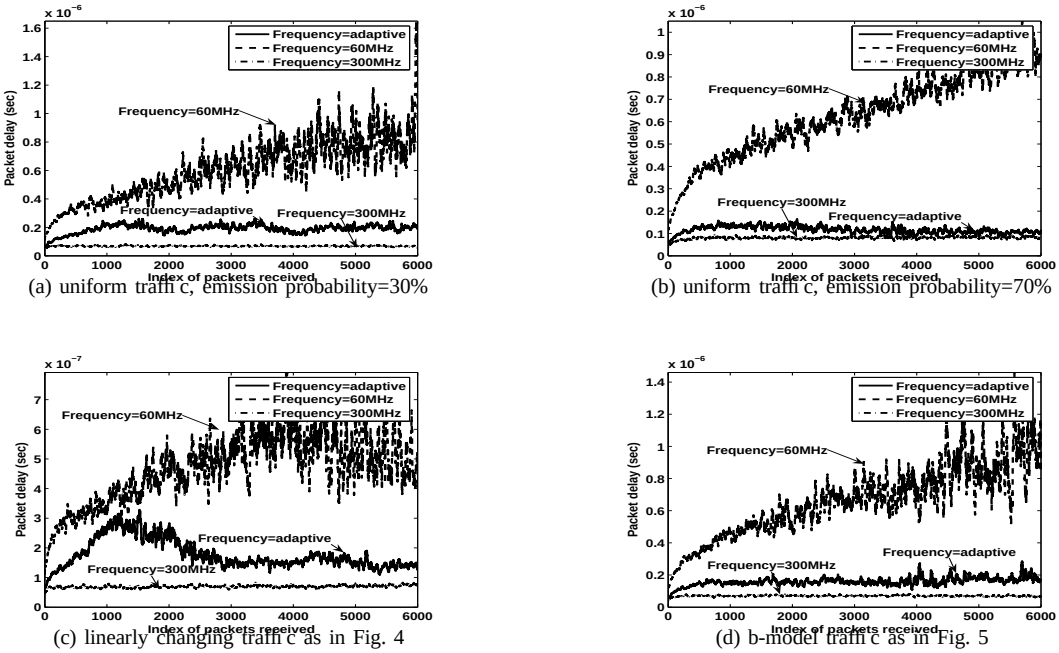
The above experiments firstly demonstrate that fixed low frequency network is not a proper choice for implementation, and in practice fixed frequency networks have to be set at a high frequency to cover the worst case. Table I compares the packet delay after transmission of 6000 packets of 60MHz frequency and adaptive control against 300MHz frequency, with their trend (stabilized or still increasing) labelled.

While adaptive control has relatively longer packet delay than fixed high frequency, it is much more power efficient than the latter, as shown in Table II, which compares average energy consumption for one packet in the simulation of 2000 cycles of adaptive control against fixed 300MHz frequency. Moreover, an adaptive control network gives the designer the possibility to make a trade-off between performance and power. If a high performance solution is desired, the target load can be set lower which will result on average in higher frequency and lower delays. For a low power solution the target load can be set higher resulting in lower frequency and higher delays.

VII. CONCLUSION

In summary we conclude that for the network to work fully loaded and with satisfying performance, there is an optimal frequency that delivers all packets in reasonable time and consumes least power. The adaptive control mechanism we have designed is able to adjust frequency and network load rapidly with a discrete range of frequency and voltage inputs, even when the voltage switching is slower than potential traffic change. As a result, the power consumption is minimized with small extra power consumed by voltage switching and the packet delay is kept to a low level. The control mechanism outperforms fixed frequency control under all kinds of traffic patterns that we studied.

Our future work will be trying to design the controller



(a) uniform traffic, emission probability=30%
(b) uniform traffic, emission probability=70%
(c) linearly changing traffic as in Fig. 4
(d) b-model traffic as in Fig. 5

Fig. 11. Packet delay for each traffic pattern under adaptive and fixed frequency control

for spatially non-uniform traffic. In that case, the measurement points have to be selected carefully. More measurement points lead to better results and less vulnerability to spatially-changing traffic patterns but increase the measurement overhead. Also, we work on reconciling adaptive frequency control with timing constraints on particular connections. For purely best effort traffic it is easy to adjust delay in order to minimize power consumption. Maximum delay requirements however impose additional constraints on the adaptive network control system.

REFERENCES

[1] N. Jha, "Low power system scheduling and synthesis. embedded tutorial," in *Proceedings of the International Conference on Computer Aided Design*, November 2001, pp. 259–263.

[2] V. Raghunathan, M. B. Srivastava, and R. K. Gupta, "A survey of techniques for energy efficient on-chip communication," in *Proceedings of the Design Automation Conference*, June 2003, pp. 900–905.

[3] L. Shang, L.-S. Peh, and N. K. Jha, "Power-efficient interconnection networks: Dynamic voltage scaling with links," *Computer Architecture Letters*, vol. 1, May 2002.

[4] A. Vitkovski, R. Haukilahti, A. Jantsch, and E. Nilsson, "Low-power and error coding for network-on-chip traffic," in *Proceedings of the IEEE NorChip Conference*, November 2004. [Online]. Available: http://www.imit.kth.se/ axel/papers/2004/NorChip-arseni-vitkowski.pdf

[5] D. Pamunuwa, J. Öberg, L.-R. Zheng, M. Millberg, A. Jantsch, and H. Tenhunen, "A study on the implementation of 2-D mesh based networks on chip in the nanoregime," *Integration - The VLSI Journal*, vol. 38, no. 1, pp. 3–17, October 2004.

[6] S. Kumar, A. Jantsch, J.-P. Soininen, M. Forsell, M. Millberg, J. Öberg, K. Tiensyrjä, and A. Hemani, "A network on chip architecture and design methodology," in *Proceedings of IEEE Computer Society Annual Symposium on VLSI*, April 2002. [Online]. Available: http://www.imit.kth.se/ axel/papers/2002/ISVLSI.pdf

[7] M. Millberg, E. Nilsson, R. Thid, S. Kumar, and A. Jantsch, "The Nostrum backbone - a communication protocol stack for networks on chip," in *Proceedings of the VLSI Design Conference*, Mumbai, India, January 2004. [Online]. Available: http://www.imit.kth.se/ axel/papers/2004/VLSI-Millberg.pdf

[8] M. Millberg, E. Nilsson, R. Thid, and A. Jantsch, "Guaranteed bandwidth using looped containers in temporally disjoint networks within the Nostrum network on chip," in *Proceedings of the Design Automation and Test Europe Conference (DATE)*, February 2004. [Online]. Available: http://www.imit.kth.se/ axel/papers/2004/DATE-Millberg.pdf

[9] E. Nilsson and J. Öberg, "Reducing peak power and latency in 2-D mesh NoCs using globally pseudochronous locally synchronous clocking," in *Proceedings of the International Conference on Hardware/Software Codesign and System Synthesis*, September 2004. [Online]. Available: http://www.imit.kth.se/ axel/papers/2004/CODES+ISSS-Erland.pdf

[10] G. Wei, J. Kim, D. Liu, S. Sidiropoulos, and M. Horowitz, "A variabla frequency parallel I/O interface with adaptive power supply regulation," *J. of Solid State Circuits*, vol. 35, no. 11, pp. 1600–1610, Nov. 2000.

[11] J. Kim and M. A. Horowitz, "Adaptive supply serial links with sub-1v operation and per-pin clock recovery," in *Proceedings of the IEEE International Solid State Circuits Conference*, 2002, pp. 268–269.

[12] E. J. Kim, K. H. Yum, G. M. Link, C. R. Das, N. Vijaykrishnan, M. Kandemir, and M. J. Irwin, "Energy optimization techniques in cluster interconnects," in *International Symposium on Low Power Electronics and Design (ISLPED'03)*, Seoul, Korea, August 2003.

[13] V. Soteriou and L.-S. Peh, "Design-space exploration of power-aware on/off interconnection networks," in *Proceedings of the 22nd International Conference on Computer Design (ICCD)*, October 2004.

[14] F. Worm, P. Ienne, P. Thiran, , and G. D. Micheli, "A robust self-calibrating transmission scheme for on-chip networks," *IEEE Transactions on very large scale integration (VLSI) systems*, vol. 12, no. 12, pp. 1360–1373, December 2004.

[15] A. J. Stratakos, "High-efficiency low-voltage DC-DC conversion for portable applications," Ph.D. dissertation, University of California, Berkeley, 1998.

[16] T. D. Burd and R. W. Brodersen, "Design issues for dynamic voltage scaling," in *Proceedings International Symposium on Low Power Electronics and Design*, Rapallo, Italy, 2000, pp. 9 – 14.

[17] W. J. Dally and J. W. Poulton, *Digitial Systems Engineering*. New York: CUP, 1998.

[18] R. Jejurikar, C. Pereira, and R. K. Gupta, "Leakage aware dynamic voltage scaling for real time embedded systems," CECS, Tech. Rep. 03-35, 2003.

[19] M. Wang, T. M. Madhyastha, N. H. Chan, S. Papadimitriou, and C. Faloutsos, "Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic," in *ICDE*, 2002. [Online]. Available: citeseer.ist.psu.edu/article/wang01data.html