ISSN 2354-0575

ĐÁNH GIÁ HIỆU NĂNG CỦA CHIP ĐA NHÂN VỚI CÁC CẤP CACHE

Hồ Khánh Lâm, Dư Đình Viên, Vũ Ngọc Hưng, Nguyễn Duy Việt, Phạm Văn Hải
Trường Đại học Sư phạm Kỹ thuật Hưng Yên

Ngày tòa soạn nhận được bài báo: 10/04/2017
Ngày phản biện đánh giá và sửa chữa: 05/05/2017
Ngày bài báo được duyệt đăng: 20/05/2017

Tóm tắt:
Sự phát triển nhanh chóng của công nghệ chip đa nhân đã làm đổi mới nhiều lĩnh vực công nghệ như điện tử-viễn thông, công nghệ thông tin. Hiệu năng của chip đa nhân với sự đưa vào các tổ chức cache đa lớp đã và đang được nhiều nhà công nghệ và nghiên cứu quan tâm. Đã có nhiều giải pháp đánh giá hiệu năng của các chip đa nhân. Bài báo của chúng tôi đưa ra một giải pháp đánh giá hiệu năng của chip đa nhân nhờ sử dụng mạng hàng đợi đóng đa lớp công việc dạng tích MCPFCQN với 05 tham số: Số lượng khách hàng, thời gian chờ đợi, thời gian đáp ứng, mức độ sử dụng và thông lượng. Kết quả cho thấy rằng khi số cấp cache tăng lên, các tham số: số lượng khách hàng, thời gian chờ đợi, mức độ sử dụng và thông lượng đều tăng lên, ngược lại, thời gian đáp ứng giảm xuống.
Từ khóa: *Chip đa nhân, MCPFCQN, hiệu năng.*

1. GIỚI THIỆU CHUNG

Chip đa xử lý (CMP) ngày nay được sử dụng trong nhiều hệ thống máy tính PC, máy tính hiệu năng cao, siêu máy tính,… Sự ra hệ thống nhớ đa cấp, trong đó có các cấp nhớ Cache trung gian tốc độ cao dựa trên công nghệ SRAM là giải pháp đem lại cuộc cách mạng trong thiết kế CMP. Ngày nay các CMP thương mại đều đưa vào các cấp cache bên trong chip (L1 và L2 cache). Tuy nhiên, xu hướng công nghệ CMP là tăng số lượng nhân, cũng làm tăng ảnh hưởng các thông số của hiệu năng như trễ truyền thông của liên kết giữa các nhân, năng lượng tiêu thụ, mức tăng tốc đạt được, mạng kết nối các nhân (OCIN) [5][6][7], công nghệ quang của kết nối OCIN [8], số luồng mà một nhân có thể xử lý, hiệu năng của cache trong CMP [9][10], các tổ chức cache [11] và các chính sách thay thế cache của CMP. Để đạt được một vài thông số hiệu năng trên cần đến các giải pháp công nghệ phức tạp cho thiết kế và chế tạo CMP. Trong bài báo này, chúng tôi đưa ra một giải pháp mô hình hóa CMP với các cấp cache sử dụng MCPFCQN để phân tích và đánh giá hiệu năng của CMP.

2. GIẢI PHÁP ĐỀ XUẤT

Mạng hàng đợi đóng đa lớp công việc [1][2] dạng tích MCPFCQN (Multiple Job Class Product Form Closed Queueing Network) là mạng hàng đợi mà trong đó không có các cửa vào và các cửa ra, thay vào đó là các liên kết hồi tiếp từ một số cửa ra của một số hàng đợi nào đó đến một số cửa vào của một số hàng đợi khác. Các lớp công việc khác nhau về xác suất định tuyến và thời gian được phục vụ. Mạng có dạng tích (PFQN) được Jackson [3] định nghĩa là mạng hàng đợi mở và đóng với các thời gian đến và các thời gian phục vụ có phân bố mũ, trong đó phân bố cân bằng là ví dụ đơn giản được xem như là mạng có dạng tích và thỏa mãn các điều kiện sau đây:

a) Nếu mạng mở, thì quá trình đến của các khách hàng từ ngoài tới nút hàng đợi là tiến trình Poisson.

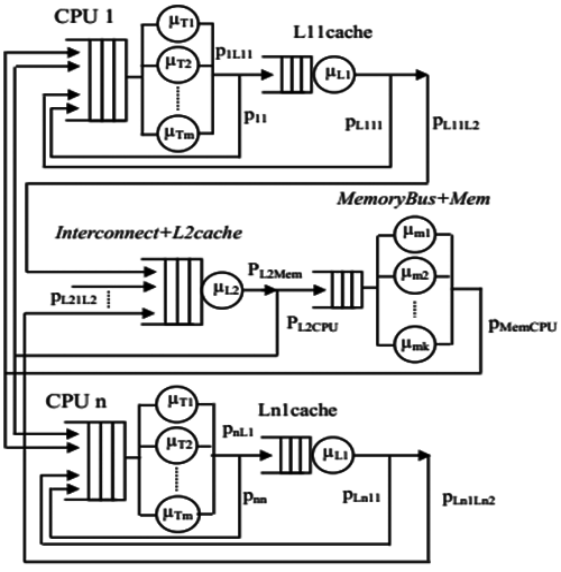
b) Tất cả thời gian phục vụ khách hàng được phân bố mũ và nguyên tắc phục vụ ở tất cả các hàng đợi là FCFS (đến trước phục vụ trước).

c) Một khách hàng hoàn thành phục vụ ở hàng đợi i hoặc là chuyển tới một số hàng mới j với xác suất P_{ij} hoặc đối với mạng mở sẽ rời khỏi hệ thống với xác suất $1 - \sum_{j=1}^{m} P_{ij}$.

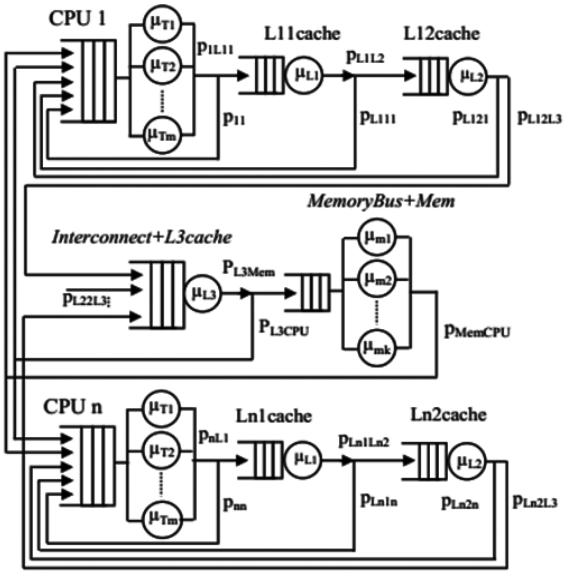
d) Hiệu suất sử dụng của tất cả các hàng đợi < 1.

e) Các PFQN có nhiều lớp công việc (khách hàng, bản tin) và có thể là hàng đợi mở đối với một số lớp công việc và hay hàng đợi đóng đối với các lớp công việc khác. Nếu là hàng đợi mở, tuân thủ theo tiêu chuẩn a.

Hiệu năng của các mạng dạng tích PFQN được phân tích và đánh giá theo hai thuật toán: thuật toán cuộn và phân tích giá trị trung bình MVA. Chúng tôi sử dụng công cụ JMT 0.9.3 dựa vào MVA để tính các thông số hiệu năng cho CMP lựa chọn.



a)



b)

Hình 1. *Mô hình MCPFCQN cho CMP đa luồng n nhân*

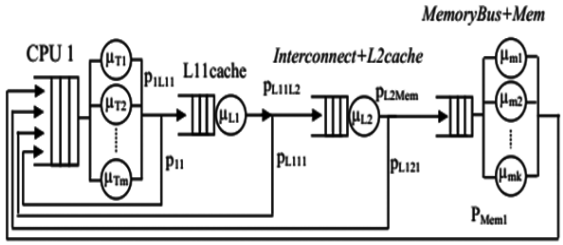
Dựa vào mô hình MCPFCQN chúng tôi để xuất mô hình mạng hàng đợi ở hình 1a cho kiến trúc CMP đa luồng n nhân, mỗi nhân có L1 và L2 cache chia sẻ chung. Hình 1b là mạng hàng đợi cho CMP với m nhân có L1 và L2 riêng, L3 chia sẻ chung. Mỗi nhân là một hàng đợi loại M/G/m-PS, với m luồng xử lý song song nên được coi là một nhân logic hay server có thời gian phục vụ trung bình là $1/\mu_{Ti}$; i = 1, 2,.., m. PS (processor sharing): là mỗi lõi CPU đưa ra nguyên tắc phục vụ của mình cho một công việc bằng việc chia sẻ nguồn tài nguyên của nó. Mạng liên kết (Interconnect) và L2 cache chia sẻ (trong hình 1a) hoặc L3 chia sẻ (trong hình 1b) là các nút quan trọng ảnh hưởng đến hiệu năng của hệ thống. Vì vậy, ở đây thiết lập mô hình cho mạng liên kết và L2 cache là một nút Interconnect+L2 cache ở hình 1a (hay Interconnect+L3 cache ở hình 1b) với thời gian phục vụ trung bình $1/\mu_{L2}$ (bao gồm thời gian truy cập L2 cache và độ trễ chuyển đổi kết nối) (hay $1/\mu_{L3}$). Bus bộ nhớ và bộ nhớ chính được đặt vào một nút MemBus+Mem với thời gian phục vụ trung bình của mỗi mô-đun là $1/\mu_{mi}$, i = 1, 2, …, k (bao gồm độ trễ bus bộ nhớ và thời gian truy cập bộ nhớ). Tất cả các cache và MemBus+Mem được mô hình hóa bằng nút hàng đợi loại M/M/1-FCFS.

Xét các thông số của CMP với ba cấp cache

Vì các lõi có cấu trúc và tài nguyên nhớ giống nhau nên trong mô hình rút gọn chỉ xét cho một nhân duy nhất và mỗi lõi chỉ thực hiện các công việc của một lớp.

Mô hình rút gọn:

• MCPFCQN rút gọn của hình 1a



Hình 2. *MCPFCQN rút gọn cho CMP đa luồng có 2 cấp cache với L2 cache chung*

Trong mô hình này mạng gồm 4 nút hàng đợi: i = 1, 2, 3, 4. Trong đó: i = 1 là nút hàng đợi CPU1; i = 2 là nút hàng đợi L11 cache; i = 3 là nút hàng đợi Interconnect+Mem; i = 4 là nút hàng đợi MemoryBus+Mem.

Đặt thời gian phục vụ trung bình tại các nút:

$$\frac{1}{\mu_1} = 0.5ns; \frac{1}{\mu_2} = 1ns; \frac{1}{\mu_3} = 2.5ns; \frac{1}{\mu_4} = 40ns$$

Đặt xác suất định tuyến tại các nút:

$p_{11} = 0.1$; $p_{12} = 0.9$; $p_{21} = 0.8$; $p_{23} = 0.2$; $p_{31} = \frac{0.8}{n}$; $p_{34} = 0.2$; $p_{41} = \frac{1}{n}$ (n-số nhân trên chip).

Tốc độ đến các nút: $v_i = \sum_{j=1}^{4} v_j p_{ij}$ với i là số nút của mạng.

Tính toán các thông số hiệu năng của CMP 2 nhân/8 luồng và L2 cache chia sẻ chung:

Áp dụng thuật toán MVA để tính toán các thông số hiệu năng, thực hiện như sau:

+ Bước 1: Khởi tạo, i = 1, 2, 3, 4

$E[N_1(0)] = E[N_2(0)] = E[N_3(0)] = E[N_4(0)] = 0$; $p_1(0/0) = 1$; $p(1/0) = 0$.

+ Bước 2: Lặp theo số lượng công việc n = 1, 2, 3,…, N

Bắt đầu từ n = 1

Bước 2.1. Thời gian đáp ứng trung bình tại các nút:

Nút 1 (CPU1):

$$E[R_1(1)] = \frac{1}{m_1\mu_1}[1 + E[N_1(1)] + \sum_{j=1}^{m_i-1}(m_i - j - 1)p_1(0/0)]$$

Nút 2 (L11 cache):

$$E[R_2(1)] = \frac{1}{\mu_2}[1 + E[N_2(1)];$$

Nút 3 (Interconnect+L2cache):

$$E[R_3(1)] = \frac{1}{\mu_3}[1 + E[N_3(1)];$$

Nút 4 (Memory Bus+Mem):

$$E[R_4(1)] = \frac{1}{\mu_4}[1 + E[N_4(1)];$$

Bước 2.2: Thông lượng toàn mạng:

$$\lambda(1) = \frac{1}{\sum_{i=1}^{4} v_i E[R_i(1)]};$$

Thông lượng của từng nút: $\lambda_i(1) = \lambda(1)v_i$;

Bước 2.3: Số lượng trung bình các công việc tại các nút mạng: $E[N_i(1)] = v_i E[R_i(1)]$;

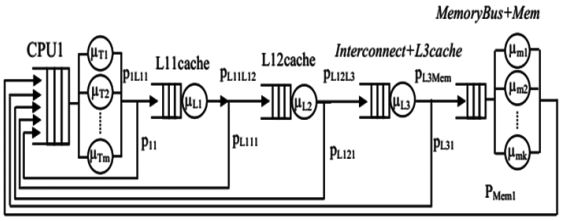
Bước 2.4: Thời gian chờ đợi trung bình các công việc tại các nút:

$$E[W_i(n)] = E[R_i(1)] - \frac{1}{\mu_i};$$

Bước 2.5: Mức độ sử dụng tại các nút: $U_i = \frac{\lambda_i}{\mu_i m_i}$.

Thực hiện lặp lại với n = 2; n = 3; ….; n = N

MCPFCQN rút gọn của hình 1b



Hình 3. *MCPFCQN rút gọn cho CMP đa luồng có 3 cấp cache với L3 cache chung*

Trong mô hình này mạng gồm 5 nút hàng đợi: i = 1, 2, 3, 4, 5.

Trong đó: i = 1 là nút hàng đợi CPU1; i = 2 là nút hàng đợi L11 cache; i = 3 là nút hàng đợi L12 cache; i = 4 là nút hàng đợi Interconnect+Mem; i = 5 là nút hàng đợi MemoryBus+Mem.

Đặt thời gian phục vụ trung bình tại các nút:

$$\frac{1}{\mu_1} = 0.5ns; \frac{1}{\mu_2} = 1ns; \frac{1}{\mu_3} = 2.5ns; \frac{1}{\mu_4} = 40ns.$$

Đặt xác suất định tuyến tại các nút:

$p_{11} = 0.1$; $p_{12} = 0.9$; $p_{21} = 0.8$; $p_{23} = 0.2$;

$p_{31} = 0.8$; $p_{34} = 0.2$; $p_{41} = \frac{0.8}{n}$; $p_{45} = 0.2$; $p_{51} = \frac{1}{n}$.

Tốc độ đến các nút: $v_i = \sum_{j=1}^{5} v_j p_{ij}$; với i là số nút của mạng.

Tính toán các thông số hiệu năng của CMP 2 nhân/8 luồng với L3 cache chia sẻ chung:

Áp dụng thuật toán MVA để tính toán các thông số hiệu năng, thực hiện như sau:

+ *Bước 1*: Khởi tạo i = 1, 2, 3, 4, 5

$E[N_1(0)] = E[N_2(0)] = E[N_3(0)] = E[N_4(0)] = E[N_5(0)] = 0$;

$p_1(0/0) = 1; p(1/0) = 0$.

+ *Bước 2*: Lặp theo số lượng công việc n = 1, 2, 3,…, N

Bắt đầu từ n = 1

Bước 2.1. Thời gian đáp ứng trung bình tại các nút:

Nút 1 (CPU1):

$$E[R_1(1)] = \frac{1}{m_1\mu_1}[1 + E[N_1(1)] + \sum_{j=1}^{m_i-2}(m_i - j - 1)p_1(0/0)]$$

Nút 2 (L11 cache):

$$E[R_2(1)] = \frac{1}{\mu_2}[1 + E[N_2(1)];$$

Nút 3 (L21 cache):

$$E[R_3(1)] = \frac{1}{\mu_3}[1 + E[N_3(1)];$$

Nút 4 (Interconnect+L2cache):

$$E[R_4(1)] = \frac{1}{\mu_4}[1 + E[N_4(1)];$$

Nút 5 (Memory Bus+Mem):

$$E[R_5(1)] = \frac{1}{\mu_5}[1 + E[N_5(1)];$$

Bước 2.2: Thông lượng toàn mạng:

$$\lambda(1) = \frac{1}{\sum_{i=1}^{5} v_i E[R_i(1)]};$$

Thông lượng của từng nút: $\lambda_i(1) = \lambda(1)v_i$;

Bước 2.3: Số lượng trung bình các công việc tại các nút:

$$E[N_i(1)] = v_i E[R_i(1)];$$

Bước 2.4: Thời gian chờ đợi trung bình các công việc tại các nút:

$$E[W_i(n)] = E[R_i(1)] - \frac{1}{\mu_i};$$

Bước 2.5: Mức độ sử dụng tại các nút: $U_i = \frac{\lambda_i}{\mu_i m_i}$.

Thực hiện lặp lại với n = 2; n = 3; ….; n = N.

3. Kết quả tính toán và đánh giá hiệu năng của kiến trúc chip đa nhân đa luồng

Chúng tôi sử dụng công cụ JMT v.0.9.3 (hoặc 0.8.0) để thực hiện mô phỏng cho các mô hình MCPFCQN ở Hình 1 theo các kịch bản:

• CMP 2 nhân/10 luồng với L2 cache chung và L3 cache chia sẻ chung

Kết quả tính toán các tham số hiệu năng: số lượng khách hàng, thời gian đợi, thời gian đáp ứng, mức độ sử dụng, thông lượng (bảng 1) ở các nhân của CPU và các cấp L1 cache, L2 cache, L3 cache của các nhân. Các kết quả ở các nhân của CPU và các cấp cache L1, L2, L3 là tương đương nhau. Do đó, ở đây chỉ trình bày kết quả của các thông số hiệu năng tại các nút Core1, L11 cache, Interconnect+L2cache; L21 cache, Interconnect+L3cache, Memory+Bus và của hệ thống trong Bảng 1.

Nhận xét:

Số lượng khách hàng (số công việc) tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì số lượng khách hàng tại nút Int+L3cache tăng 42% và tại nút MemBus+Mem giảm 53% so với số lượng khách hàng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thời gian chờ đợi tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thời gian đáp ứng tại nút Int+L3cache giảm 59% và tại nút MemBus+Mem giảm 99% so với thời gian chờ đợi tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thời gian đáp ứng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì số thời gian đáp ứng tại nút Int+L3cache tăng 72% và tại nút MemBus+Mem giảm 52% so với thời gian đáp ứng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Mức độ sử dụng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì mức độ sử dụng tại nút Int+L3cache tăng 20% và tại nút MemBus+Mem giảm 59% so với mức độ sử dụng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thông lượng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thông lượng tại nút Int+L3cache giảm 38% và tại nút MemBus+Mem giảm 36% so với thông lượng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache.

Bảng 1. *Giá trị trung bình các thông số hiệu năng của CMP 2 nhân 8 luồng/nhân*

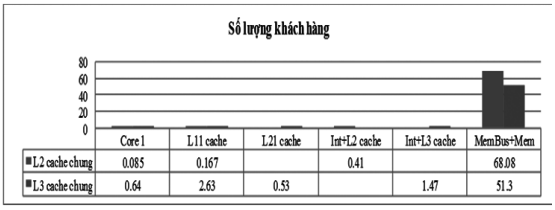
	Số lượng khách hàng		Thời gian chờ đợi (ns)		Thời gian đáp ứng (ns)		Mức độ sử dụng		Thông lượng	
	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung
Core 1	0,157	1,08	0,499	1,09	0,5	1,1	0,09	0,55	0,34	1,09
L11 cache	0,41	37,4	1,3	34,3	1,45	38,12	0,31	0,98	0,31	0,98
L12 cache		0,86		0,88		4,9		0,49		0,19
Int+L2 cache	0,409		1,3		3,63		0,31		0,13	
Int+L3 cache		0,58		0,59		8,2		0,39		0,08
MemBus+Bus	67,952	1,5	236,57	1,53	3288,6	106,81	1	0,63	0,025	0,016
System			268,46		85,48				0,31	0,98

• CMP 4 nhân/10 luồng với L2 cache chung và L3 cache chung

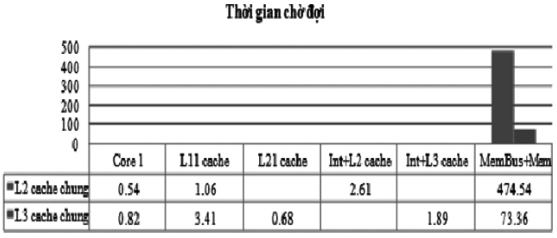
Kết quả mô phỏng cho ở Bảng 2.

Bảng 2. *Giá trị trung bình các thông số hiệu năng của CMP 4 nhân 10 luồng/nhân*

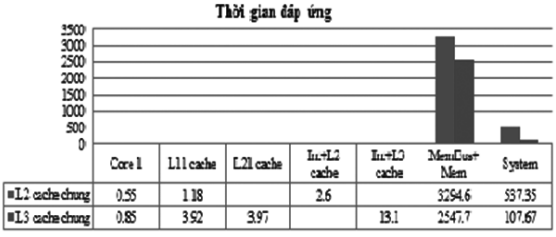
	Số lượng khách hàng		Thời gian chờ đợi (ns)		Thời gian đáp ứng (ns)		Mức độ sử dụng		Thông lượng	
	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung
Core 1	0,085	0,64	0,54	0,82	0,55	0,85	0,08	0,42	0,17	0,83
L11 cache	0,167	2,63	1,06	3,41	1,18	3,92	1,57	0,75	0,16	0,75
L21 cache		0,53		0,68		3,97		0,37		0,15
Int+L2 cache	0,41		2,61		2,6		0,31		0,125	
Int+L3 cache		1,47		1,89		13,1		0,62		0,124
MemBus+Bus	68,08	51,3	474,54	73,36	3294,6	2547,7	1	0,99	0,025	0,024
System					537,35	107,67			0,16	0,77



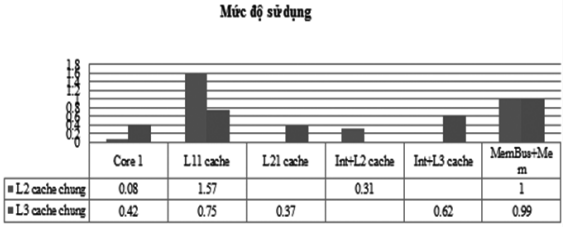
Hình 4a. *Giá trị trung bình của số lượng khách hàng ở các nút của CMP 4 nhân/10 luồng với L2 cache chung và L3 cache chung*



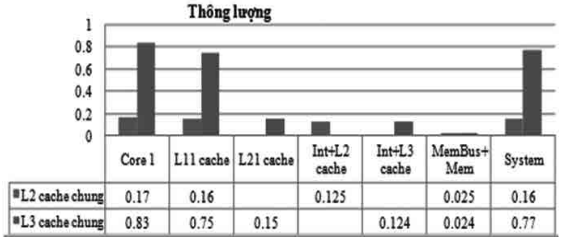
Hình 4b. *Giá trị trung bình của thời gian chờ đợi (ns) ở các nút của CMP 4 nhân/10 luồng với L2 cache chung và L3 cache chung*



Hình 4c. *Giá trị trung bình của thời gian đáp ứng (ns) ở các nút của CMP 4 nhân/10 luồng với L2 cache chung và L3 cache chung*



Hình 4d. *Giá trị trung bình của mức độ sử dụng ở các nút của CMP 4 nhân/10 luồng với L2 cache chung và L3 cache chung*



Hình 4e. *Giá trị trung bình của thông lượng ở các nút của CMP 4 nhân/10 luồng với L2 cache chung và L3 cache chung*

Nhận xét:

Số lượng khách hàng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì số lượng khách hàng tại nút Int+L3cache tăng 72% và tại nút MemBus+Mem giảm 25% so với số lượng khách hàng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thời gian tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thời gian chờ đợi tại nút Int+L3cache giảm 26% và tại nút MemBus+Mem giảm 84% so với thời gian chờ đợi tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thời gian đáp ứng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thời gian đáp ứng tại nút Int+L3cache tăng 80% và tại nút MemBus+Mem giảm 23% so với thời gian đáp ứng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Mức độ sử dụng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì mức độ sử dụng tại nút Int+L3cache tăng 52% và tại nút MemBus+Mem giảm 1% so với mức độ sử dụng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thông lượng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thông lượng của cả hệ thống tăng 80% so với thông lượng của cả hệ thống của CMP có 2 cấp cache.

- **CMP 8 nhân/10 luồng với L2 cache chung và L3 cache chung**

Bảng 3. *Giá trị trung bình các thông số hiệu năng của CMP 8 nhân/10 luồng*

	Số lượng khách hàng		Thời gian chờ đợi (ns)		Thời gian đáp ứng (ns)		Mức độ sử dụng		Thông lượng	
	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung	L2 cache chung	L3 cache chung
Core 1	0.04	0,25	0,52	0,63	0,52	0,64	0,04	0,22	0,087	0,43
L11 cache	0.07	0,58	0,97	1,47	1,08	1,64	0,08	0,39	0,078	0,39

L21 cache		0,22		0,55		3,09		0,19		0,078
Int+L2 cache	0,4		5,22		3,63		0,31		0,12	
Int+L3 cache		1,49		3,83		13,35		0,62		0,125
MemBus+Bus	66.26	60,48	924,08	169,7	3206,9	2952,9	1	0,99	0,025	0,025
System					1074,8	214,66			0,078	0,39

Nhận xét:

Số khách hàng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì Số khách hàng tại nút Int+L3cache tăng 73% và tại nút MemBus+Mem giảm 9% so với số khách hàng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thời gian chờ đợi tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thời gian chờ đợi tại nút Int+L3cache giảm 28% và tại nút MemBus+Mem giảm 82% so với thời gian chờ đợi tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thời gian đáp ứng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thời gian đáp ứng tại nút Int+L3cache tăng 72% và tại nút MemBus+Mem giảm 8%, cả hệ thống giảm 80% so với thời gian đáp ứng tại nút Int+L2cache và MemBus+Mem, hệ thống của CMP có 2 cấp cache. Mức độ sử dụng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì mức độ sử dụng tại nút Int+L3cache tăng 50% và tại nút MemBus+Mem giảm 1% so với mức độ sử dụng tại nút Int+L2cache và MemBus+Mem của CMP có 2 cấp cache. Thông lượng tại các nút chia sẻ là rất lớn, với CMP có 3 cấp cache thì thông lượng của cả hệ thống tăng 400% so với thông lượng của cả hệ thống của CMP có 2 cấp cache.

Kết quả mô phỏng cho thấy rằng: Đối với chip đa nhân có 3 cấp cache, tại các nút Int+L3cache và MemBus+Mem có số lượng khách hàng, thời gian chờ đợi, thời gian đáp ứng và mức độ sử dụng tăng lên, nhưng thời gian chờ đợi lại giảm nhiều so với chip đa nhân có 2 cấp cache. Thông lượng của chip đa nhân có 3 cấp cache cũng lớn hơn thông lượng của chip đa nhân có 2 cấp cache. Điều này chứng tỏ rằng, với chip đa nhân có 3 cấp cache làm giảm đáng kể độ trễ và thời gian truy nhập bộ nhớ, do đó giảm nghẽn cổ chai tại các cấp cache chia sẻ và tăng hiệu năng của bộ xử lý.

Với các dữ liệu mặc định: L1 hit time = 1ns, L2 hit time = 2.5ns, L3 hit time = 5ns, MAT = 40ns, L1 miss rate = 0.2, L2 miss rate = 0.2, L3 miss rate = 0.2, xác định được thời gian truy nhập trung bình bộ nhớ (AMAT), mức tăng tốc (SP) của từng kiến trúc, từ đó đánh giá được hiệu năng của chip đa nhân có 3 cấp cache so với chip đa nhân có 2 cấp cache:

Đối với chip đa nhân có 3 cấp cache (L1, L2, L3):

+ Thời gian truy nhập trung bình bộ nhớ chính: MAT = 40ns

+ Thời gian truy nhập trung bình bộ nhớ được tính bằng:

AMAT = L1 hit time + (L1 miss rate) x (L2 hit time + (L2 miss rate) x (L3 hit time) + (L3 miss rate) x (MAT)))

AMAT = 1ns + (0.2)(2.5ns + (0.2)(5ns + (0.2)(40ns))) = 2.02ns

+ Mức tăng tốc của hệ thống:

$$SP = \frac{MAT}{AMAT} = \frac{40}{2.02} = 19.8$$

Đối với chip đa nhân có 2 cấp cache (L1, L2):

+ Thời gian truy nhập trung bình bộ nhớ chính: MAT = 40ns

+ Thời gian truy nhập trung bình bộ nhớ:

AMAT = L1hit time + (L1miss rate) x (L2 hit time + (L2 miss rate) x (MAT))

AMAT = 1ns + (0.2)(2.5ns + (0.2)(40ns)) = 3.1ns

+ Mức tăng tốc của hệ thống:

$$SP = \frac{MAT}{AMAT} = \frac{40}{3.1} = 12.8$$

CMP có 3 cấp cache thì thời gian truy nhập trung bình bộ nhớ giảm đi: 3,1 – 2,02 = 1,08ns, mức tăng tốc của hệ thống tăng 1,5 lần so với chip đa nhân có 2 cấp cache. Có thể thấy rằng, với kiến trúc chip đa nhân có 3 cấp cache với L3 cache chia sẻ cho kết quả khả quan, giảm được thời gian trung bình truy nhập bộ nhớ, giảm nghẽn cổ chai tại các nút chia sẻ, do đó nâng cao được hiệu năng của CMP.

4. KẾT LUẬN

Nghiên cứu về kiến trúc CMP và ảnh hưởng tổ chức cache trong kiến trúc chip đa nhân đã được thực hiện trong thời gian dài, những vẫn được quan tâm do tầm quan trọng và sự ảnh hưởng của nó đối với hiệu năng của hệ thống máy tính. Mô hình hóa CMP bằng MCPFCQN là giải pháp hiệu quả cho phép thực hiện mô phỏng và đánh giá hiệu năng của bất cứ loại CMP nào mong muốn và nó là công cụ tốt để tham khảo cho tư vấn thiết kế hoặc sử dụng CMP. Giải pháp trình bày ở đây đã đưa ra mô hình rút gọn, xây dựng các biểu thức tính các tham số hiệu năng và sau đó tính toán các tham số hiệu năng. Kết quả tính toán cho thấy rằng khi số cấp cache tăng lên, các tham số: số lượng khách hàng, thời

gian chờ đợi, mức độ sử dụng và thông lượng đều tăng lên, ngược lại, thời gian đáp ứng giảm xuống. Lưu ý rằng giải pháp chưa cân nhắc các tham công nghệ khác của CMP như cấu hình liên kết các nút (OCIN), dung lượng các cấp cache, các thuật toán thay thế cache, số lượng nhiều nhân, công suất tiêu thụ hay lượng tán nhiệt. Đó là những thông số cần phải tính đến trong phân tích ảnh hưởng đến hiệu năng của CMP với hàng trăm, hàng nghìn nhân cho tương lai phát triển của công nghệ CMP.

Tài liệu tham khảo

[1]. J. Virtamo, "*Queueing Theory / Probability Theory*", www.netlab.hut.fi/opetus/.

[2]. Kiran M Rege, "*Multi-class Queueing Models for Performance Analysis of Computer Systems*", December 1990, Volume 15, Issue 4, pp. 355–363. DOI: 10.1007/BF02811331.

[3]. Jackson, R. R. P. (1995). "*Book Review: Queueing Networks and Product Forms: A Systems Approach*". IMA Journal of Management Mathematics. 6 (4): 382–384. doi:10.1093/imaman/6.4.382.

[4]. Daniel Sanchez, George Michelogiannakis, and Chitistos Kozyrakis, "*An Analysis of On-Chip Interconnection Networks for Large-Scale Chip Multiprocessors*". Stanford University. ACM Transactions on Architecture and Code Optimization, Vol. 7, No. 1, Article 4, Publication date: April 2010.

[5]. David Wentzlaff et all, "*On – chip Intercinnection Architecture of the Title Processor*". 0272-1732/07/$20.00 G 2007 IEEE. Published by the IEEE Computer Society. Authorized licensed use limited to: The University of Toronto. Downloaded on January 4, 2010 at 12:39 from IEEE Xplore.

[6]. D. N. Jayasimha, Bilal Zafar, Yatin Hoskote,"*On-Chip Interconnection Networks: Why They are Different and How to Compare Them*".

[7]. Jesús Camacho Villanueva et all, "*A Performance Evaluation of 2D-Mesh, Ring, and Crossbar Interconnects for Chip Multi-Processors*". NoCArc '09, December 12, 2009, New York City, New York, USA Copyright © 2009 ACM 978-1-60558-774-5.

[8]. B. Krishna Priya, Amit D. Joshi, N. Ramasubramanian, "*A Survey on Performance of On-Chip Cache for Multi-core Architecture*", Pondicherry, India — August 25 - 26,2016ISBN:978-1-4503-4756-3.

[9]. Jie Tao, Marcel Kunze, Fabian Nowak, Rainer Buchty, Wolfgang Karl, "*Performance Advantage of Reconfigurable Cache Design on Multicore Processor Systems*", Int J Parallel Prog (2008) 36:347–360. DOI 10.1007/s10766-008-0075-4.

[10]. Zvika Guz, Idit Keidar, Avinoam Kolodny, Uri C. Weiser, "*Nahalal: Cache Organization for Chip Multiprocessors*", Manuscript submitted: 24-Apr-2007. Manuscript accepted: 23-May-2007. Final manuscript received: 29-May-2007.

[11]. Muhammad Ali Ismail, "*Performance Behavior Analysis of the Present 3-Level Cache System for Multi-Core Systems using Queuing Modeling*", International Conference on Latest Computational Technologies (ICLCT'2012) March 17-18, 2012 Bangkok.

EVALUATING PERFORMANCE OF CHIP MULTI-CORE WITH CACHE LEVEL

Abstract:

Chip multi-core are applied widely in hight performance computer systems and suppercomputers. The performance of CMPs with applications of cache multi-level structures is interested in many researchers. There are many solutions used to evaluate the performance of MCP. For this objective, the paper proposes an approach that uses MCPFCQN. The performance of CMP is characterised by 05 parameters: number of jobs, waiting time, response time, utilization and capacity. The results show that when the number of caches increases, number of jobs, waiting time, utilization and capacity are increased too, but only response time is deacreased.

Keywords: *Chip multi-core, Multiple Job Class Product Form Closed Queueing Network (MCPFCQN), performance.*