# Performance and Network Power Evaluation of Tightly Mixed SRAM NUCA for 3D Multi-core Network on Chips

Yuang Zhang[1, 2], Li Li[1], Zhonghai Lu[2], Axel Jantsch[2], Yuxiang Fu[1], Minglun Gao[1]

[1]Institute of VLSI Design,
Key Laboratory of Advanced Photonic and Electronic
Materials, Nanjing University
Nanjing, China
Email: {zhangyuang, lili, gaominglun} @ nju.edu.cn

[2]Department of Electronic, Computer and Software Systems,
School of Information and Communication Technology,
Royal Institute of Technology
Stockholm, Sweden
{yazhang, zhonghai, axel} @ kth.se

***Abstract*—Last level cache (LLC) is crucial for the performance of chip multiprocessors (CMPs), while power is a significant design concern for 3D CMPs. In this paper, we focus on the SRAM-based Non-Uniform Cache Architecture (NUCA) for 3D Multi-core Network-on-Chip (McNoC) systems. A tightly mixed SRAM NUCA for 3D mesh NoC is presented and analyzed. We evaluate the performance and network power with benchmarks based on a full system simulation framework. Experiment results on 16-core 3D NoC systems show that the tightly mixed NUCA could provide up to 31.71% and on average 5.95% performance improvement compared to a base 3D NUCA scheme. The tightly mixed 3D NUCA NoC can reduce network power consumption in 1.07%-15.74% and 9.64% on average compared to a baseline 3D NoCs. Our analysis and experimental results provide a guideline to design efficient 3D NoCs with stacking NUCA.**

***Keywords—3D Chip; NoC; NUCA; Multi-core***

## I. Introduction

Recently, the performance of microprocessor is getting enhanced by increasing the number of cores on a chip. 3D die-stacking technology is envisioned as a solution for future chip-multiprocessor (CMP) design. 3D CMPs support stacking memory layers atop processor layers connected by massive vertical TSVs, which are believed to mitigate the “Memory Wall” problem [1]. 3D CMP research has drawn great intentions from academic community and semiconductor industry. Several 3D CMP prototypes are emerging. Kim et al. [2] demonstrate a 3D multi-core system with one 64-core layer and one 256KB SRAM layer in 130nm technology. The processing foundries are Global Foundries device technology and Tezzaron TSV/bonding technology. Fick et al. [3] [4] propose a low power 64-core system that is called Centip3De. Centip3De has two stacked dies with a layer of 64 ARM M3 near-threshold cores and a cache layer. In a more recent work [5], Centip3De is extended to be a reconfigurable 7-layer 3D multi-core system, with 128 cores and 256 MB of DRAMs. Wordeman et al. [6] present a prototype of a 3D system with a memory layer and a logic layer connected by TSVs.

As the number of cores in CMP systems increases, Network on Chips (NoCs) [7] can offer high communication bandwidth, high flexibility and simple modular network structure with great ability of fault tolerance. Hence NoC is a promising interconnect backbone for future 3D CMPs. Several works study the interconnect topics on 3D NoC [8] [9] [10]. In this paper, we focus on the last level cache (LLC) architecture in 3D NoC context.

The on chip cache hierarchy takes an important role to feed multiple hungry cores, which not only affects the overall performance but also impacts the network communication and hence the network power consumption. For a CMP, the L1 caches are usually private and associated with each core. The L2 caches are shared by all cores to efficiently utilize the total capacity of the L2 caches, which act as the LLCs. The on chip shared LLCs are usually organized as Non-Uniform Cache Architecture (NUCA) [11] [12]. NUCA allows nearer cache banks to have lower access latencies than further banks. The shared NUCA may heavily affect the performance of a multi-core system [13]. 3D NoC provides a design exploration space of locating the banks of NUCA in different layers.

In this paper, we study the NUCA organization for 3D NoCs. A tightly coupled SRAM NUCA is presented for 3D NoCs. We evaluated the performance and network power aspects of the tightly mixed 3D NUCA NoC based on a full system simulation system. The experiments show that the presented tightly mixed 3D NUCA organization has a performance improvement up to 31.71% and on average 5.95% compared to the base 3D NUCA NoC. The network power is reduced up to 15.74% and 9.64% on average for all the test cases.

The rest of this paper is organized as follows. Section 2 discusses related works. In section 3, we present and analyze the tightly mixed NUCA for the 3D NoC. Section 4 provides simulation methodology and experiment results. Finally, we conclude this paper in section 5.

## II. Related Work

There are several works on directly stacking caches on top of processor layer for 3D CMPs. An illustration of the cache-stacking 3D CMP is shown in Fig. 1.

Black et al. [14] study small scale 3D CMPs with stacked SRAM and DRAM L2 caches on top of processors. The

978-1-4799-3432-4/14/$31.00 ©2014 IEEE

authors present three 3D stacking structures: 1) 8 MB SRAM cache on top of the base line processor layer; 2) 32 MB DRAM cache on the bare processor layer with no SRAMs; and 3) 64 MB DRAM atop the processor layer which contains 4 MB SRAM. In architecture 3), the 4 MB SRAM is used to store the tags for DRAM caches. However, the LLC architecture is not discussed in deep, and the scale of the CMP is small.

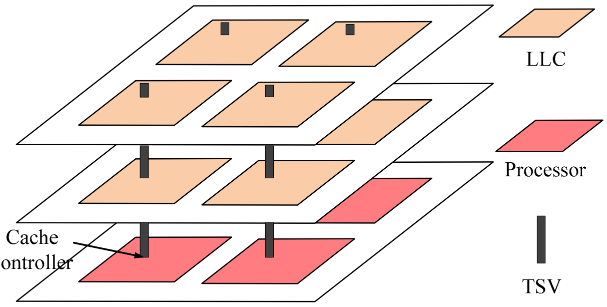

Fig. 1. An illustration of 3D cache-stacking CMP.

Xu et al. [15] evaluates stacking DRAM as LLC for 3D NoCs which consist of 2-layer, 16-core. The experiment results show that the average cache hit latency in DRAM LLC is 27.97% worse with a 25.78% power reduction compared to SRAM design. However, DRAM is usually in different process from processors. DRAM caches are in a separate layer. There is little design space for locating 3D DRAM caches. Li et al. [16] study the L2 cache design and management in 3D chip multiprocessors. The proposed 3D CMP architecture places CPUs on several layers with the remaining space filled with L2 cache banks. However, the authors use a time division multiple access (dTDMA) bus and assume that the density of TSVs may be low and allocate multiple processors with one group of TSVs. In this paper, we discuss the LLC architecture in a full mesh 3D NoC and each processor is associated with a group of TSVs. We also evaluate network power consumption which is not discussed in [16].

## III. NUCA FOR 3D MCNOC

Caches have been playing an important role in bridging the performance gap between high-speed processors and slow off-chip main memories [17]. In 3D NoC, there are several design options for the location of processors and SRAM LLCs. We analyze the memory access latency of the 3D NoCs and present our tightly mixed LLC architecture.

The average data access latency can be presented in equation (1).

$$T_{avg} = h_{L1} \times T_{L1} + m_{L1} \times \left( h_{L2} \times T_{L2} + m_{L2} \times T_{MEM} \right) \quad (1)$$

In (1), the L1 cache performance, including L1 hit latency ($T_{L1}$) and L1 hit rate ($h_{L1}$) and miss rate ($m_{L1}$), is determined by the L1 cache parameters, e.g. L1 cache size, L1 cache line size, L1 cache associativity, etc. Similarly, L2 cache hit rate ($h_{L2}$) is determined by L2 cache parameter. However, L2 hit latency ($T_{L2}$) for 3D NoC varies with the distance between the processors and the L2 caches. In 3D NoC, SRAM L2 caches can be in the same layers with cores or in separate layers. Hence, we present a tightly mixed NUCA for 3D mesh NoC which is shown in Fig. 2.

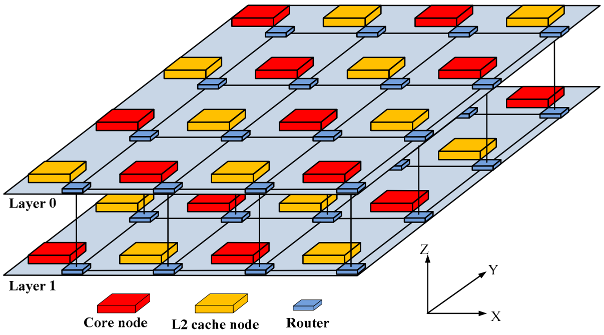

Fig. 2. Tightly mixed NUCA for 3D NoC.

In Fig. 2, the processor nodes and the L2 cache nodes are mixed. The processors consume the majority of the overall power consumption of the 3D CMP. Hence, compared to stacking processors on top of each other, the mixed NUCA 3D CMP can mitigate the potential rise of thermal issues. The base 3D NoC is shown in Fig. 3, in which the processor nodes and L2 cache nodes are in separate layers.

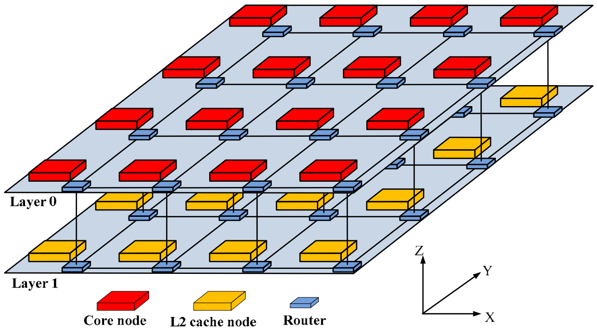

Fig. 3. Base stacking NUCA for 3D NoC.

We assume the link latency between two routers is 2 cycles and the latency between the router and the processing node (core node or L2 cache node) is 1.5 cycles. The probability that the cores access to all the 16 L2 caches is equal. The numbering of the nodes is shown in Fig. 4.

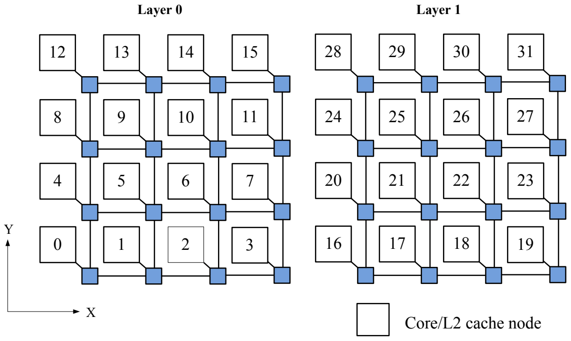

Fig. 4. Numbering of 3D NoC.

The routing algorithm we use is static XYZ. The packet is routed one dimension at a time until it reaches the same coordinate as the destination in that dimension [18]. The packets route by the X direction first, then by the Y direction and by the Z direction at last. For example, if a core at node 0 (the coordinate is 0, 0, 0) sends a read request to a L2 cache at node 31 (the coordinate is 3, 3, 2), the request goes along the following route 0-1-2-3-7-11-15-31, and the total zero-load latency is 1.5×2+2×7=17 cycles. The zero-load latency of a network is the latency where packets traverse the network without any contention. In table 1, we calculate the total zero-load latency of all the core nodes to any other cache node.

TABLE I. COMPARISON OF ZERO-LOAD LATENCY ON TIGHTLY MIXED NUCA (TM NUCA) AND BASE NUCA FOR 3D NOC IN CYCLES

| | TM NUCA | Base NUCA | | TM NUCA | Base NUCA |
|---|---|---|---|---|---|
| CPU0 | 160 | 176 | CPU8 | 144 | 160 |
| CPU1 | 144 | 160 | CPU9 | 160 | 144 |
| CPU2 | 128 | 160 | CPU10 | 144 | 144 |
| CPU3 | 144 | 176 | CPU11 | 128 | 160 |
| CPU4 | 144 | 160 | CPU12 | 128 | 176 |
| CPU5 | 128 | 144 | CPU13 | 144 | 160 |
| CPU6 | 144 | 144 | CPU14 | 160 | 160 |
| CPU7 | 160 | 160 | CPU15 | 144 | 176 |
| **Total** | 2304 | 2560 | **Average** | 144 | 160 |

From table 1, we can find that the tightly mixed NUCA (TM NUCA) has $(2560-2304)/2304=11.1\%$ less communication latency. According to Amdahl's law [19], the speedup can be calculated in (2):

$$Speedup_{overall}=\frac{Execution\_time_{old}}{Execution\_time_{new}}$$
$$=\frac{1}{(1-Fraction_{enhanced})+Fraction_{enhanced}/Speedup_{enhanced}} \quad (2)$$

If the latency caused by accessing L2 cache in networks takes up to 10% of total processing time, then the speedup is $\frac{1}{(1-0.1)+0.1/1.111}=1.01$.

If the L2 cache accessing latency takes a larger proportion, the speedup will further increase. Hence, the tightly mixed LLC should outperform the base stacking case. However, for real life applications, the data accessing pattern may not follow the uniform distribution pattern. In Section 4, we evaluate the performance, as well as the network power of the presented tightly mixed SRAM NUCA for 3D NoCs.

## IV. EXPERIMENTAL EVALUATION

In this section, we first briefly describe our simulation based experimental methodology and then present and analyze the evaluation results in terms of performance and network power.

### A. Methodology

To evaluate the performance benefit of the tightly mixed SRAM LLC for 3D NoC, we run the simulation using the GEM5 [20] full system simulator with the standard PARSEC [21] benchmark. The system parameters are listed in table 2.

There are two levels of on-chip caches in the 3D NoC memory hierarchy. The L1 instruction and data caches are private and associated with each core. The L2 caches are the shared LLCs and organized as tightly mixed NUCA. We use Orion [22] to obtain the network power consumption.

TABLE II. MULTI-CORE NOC SYSTEM CONFIGURATION PARAMETERS

| | |
|---|---|
| Processors | 16 ALPHA cores, 2GHz, Timing Simple CPU |
| OS | tsunami |
| L1 cache | 16 KB per core, split I/D, 2 way associative<br>3 cycles latency, 64 byte lines |
| L2 cache | 16 MB, 16 banks, 16 way associative<br>15 cycles latency, 64 byte lines, inclusive |
| Memory | 512 MB, 4 banks, 300 cycles latency |
| Directory | L1 tag replication, MESI protocol<br>16 banks interleaved |
| Network | 4×4×2 3D mesh topology<br>5-stage pipeline routers, 1-cycle link latency |

### B. Experiment results and analysis

The performance improvement of tightly mixed NUCA 3D NoC compared to that of the base 3D NUCA is shown in Fig. 5.

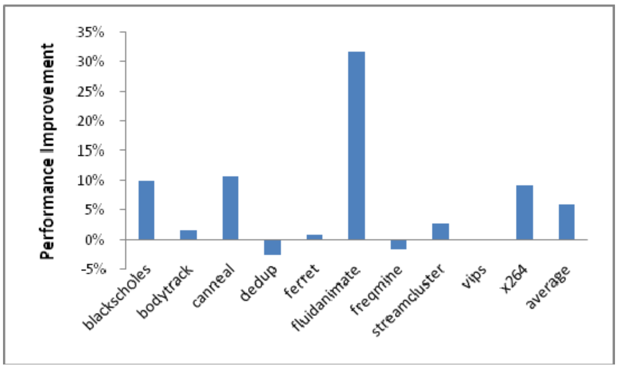

Fig. 5. The performance enhancement of tightly mixed NUCA 3D NoC.

From Fig. 5, we can observe that the performance for the tightly mixed NUCA outperforms the base case for 8 out of 10 benchmarks. And for the left 2 benchmarks, the performance has slight decline (2.6% for dedup and 1.48% for freqmine). The fluidanimate demonstrates the most significant performance improvement which is 31.71%. The performance is increased by 5.95% on average.

Fig. 6 shows the network power reduction of the tightly mixed NUCA 3D NoC over the base case.

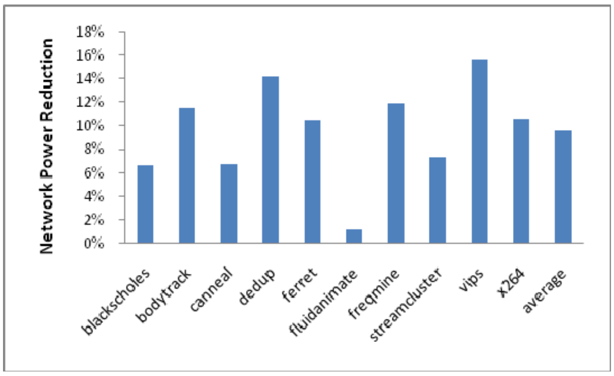

Fig. 6. The reduction of network power consumption.

The accesses to L2 caches need to go through the routers and links of the network. The network latency and power consumption are determined by the hop counts between the cores and the shared LLCs. We can find from Fig. 6 that the network power consumptions for all the benchmarks have a saving in the range from 1.07% to 15.74%. The average power reduction is 9.64%.

## V. Conclusions

In this paper, we present a tightly mixed SRAM NUCA for 3D multi-core NoC. We quantitatively analyze the performance benefits for the presented NUCA scheme. Meanwhile, we evaluate the performance and network power of both the tightly mixed SRAM NUCA and the base 3D NoC case based on an integrated simulation framework. The experiments results indicate that when the distance between processors and caches decreases, 3D NoC shows compatible performance and network power gain. The tightly mixed SRAM NUCA has up to 31.71% performance improvement and saves the network power as much as 15.74%. The results of this paper emphasize the importance of considering 3D technology in placement of LLCs for future NoCs.

## VI. ACKNOWLEDGMENTS

This work is supported by the National Nature Science Foundation of China under Grant No. 61176024, 61006018, 61370040 and 61376075; Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20120091110029; IUR Cooperative Innovation Foundation-Prospective Research Project of Jiangsu Province under Grant No.BY2013072-05; A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The authors would like to thank the anonymous reviewers for their feedback and suggestions.

## REFERENCES

[1] Jacob, P.; Zia, A.; Erdogan, O.; Belemjian, P.; Kim, J.-W.; Chu, M.; Kraft, R.; McDonald, J. & Bernstein, K., "Mitigating memory wall effects in high-clock-rate and multicore CMOS 3-D processor memory stacks", *Proceedings of the IEEE,* volume 97, issue 1, 2009, pp. 108 -122.

[2] Kim, D. H.; Athikulwongse, K.; Healy, M.; Hossain, M.; Jung, M.; Khorosh, I.; Kumar, G.; Lee, Y.-J.; Lewis, D.; Lin, T.-W.; Liu, C.; Panth, S.; Pathak, M.; Ren, M.; Shen, G.; Song, T.; Woo, D. H.; Zhao, X.; Kim, J.; Choi, H.; Loh, G.; Lee, H.-H. & Lim, S. K., "3D-MAPS: 3D massively parallel processor with stacked memory", *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, Feb. 2012, pp. 188 -190.

[3] Fick, D.; Dreslinski, R.; Giridhar, B.; Kim, G.; Seo, S.; Fojtik, M.; Satpathy, S.; Lee, Y.; Kim, D.; Liu, N.; Wieckowski, M.; Chen, G.; Mudge, T.; Sylvester, D. & Blaauw, D., "Centip3De: A 3930 DMIPS/W configurable near-threshold 3D stacked system with 64 ARM Cortex-M3 cores", *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, Feb. 2012, pp. 190-192.

[4] Fick, D.; Dreslinski, R.; Giridhar, B.; Kim, G.; Seo, S.; Fojtik, M.; Satpathy, S.; Lee, Y.; Kim, D.; Liu, N.; Wieckowski, M.; Chen, G.; Mudge, T.; Blaauw, D. & Sylvester, D., "Centip3De: A cluster-based NTC architecture with 64 ARM Cortex-M3 cores in 3D stacked 130 nm CMOS", *IEEE Journal of Solid-State Circuits*, volume 48, issue 1, 2013, pp. 104 -117.

[5] Dreslinski, R. G.; Fick, D.; Giridhar, B.; Kim, G.; Seo, S.; Fojtik, M.; Satpathy, S.; Lee, Y.; Kim, D.; Liu, N.; Wieckowski, M.; Chen, G.; Sylvester, D.; Blaauw, D. & Mudge, T., "Centip3De: A 64-Core, 3D stacked near-threshold system", *IEEE Micro*, volume 33, issue 2, 2013, pp. 8-16.

[6] Wordeman, M.; Silberman, J.; Maier, G. & Scheuermann, M., "A 3D system prototype of an eDRAM cache stacked over processor-like logic using through-silicon vias", *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, Feb. 2012, pp. 186 -187.

[7] Kumar, S.; Jantsch, A.; Soininen, J.-P.; Forsell, M.; Millberg, M.; Oberg, J.; Tiensyrja, K. & Hemani, A., "A network on chip architecture and design methodology", *IEEE Computer Society Annual Symposium on VLSI*, April 2002, pp. 105-112.

[8] Pavlidis, V. & Friedman, E., "3-D topologies for Networks-on-Chip", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, volume 15, issue 10, 2007, pp. 1081-1090.

[9] V. F. Pavlidis and E. G. Friedman, "3-D topologies for networks-on-chip", *Proc. IEEE Int. SOC Conf.*, 2006, pp. 285-288.

[10] Akbari, S.; Shafiee, A.; Fathy, M. & Berangi, R., "AFRA: A low cost high performance reliable routing for 3D mesh NoCs", *Design, Automation Test in Europe Conference Exhibition*, March 2012, pp. 332 -337.

[11] Kim, C.; Burger, D. & Keckler, S., "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches", *Proceedings of the 10th international conference on Architectural support for programming languages and operating systems*, Dec. 2002, pp. 211-222.

[12] Huh, J.; Kim, C.; Shafi, H.; Zhang, L.; Burger, D. & Keckler, S, "A NUCA substrate for flexible CMP cache sharing", *IEEE Transactions on Parallel and Distributed Systems*, volume 18, issue 8, 2007, pp. 1028-1040.

[13] B. M. Rogers, A. Krishna, G. B. Bell, K. Vu, X. Jiang, and Y. Solihin., "Scaling the bandwidth wall: challenges in and avenues for cmp scaling", *Proceedings of the 36th annual international symposium on computer architecture*, June 2009, pp. 371-382.

[14] Black, B.; Annavaram, M.; Brekelbaum, N.; DeVale, J.; Jiang, L.; Loh, G. H.; McCaule, D.; Morrow, P.; Nelson, D. W.; Pantuso, D.; Reed, P.; Rupley, J.; Shankar, S.; Shen, J. & Webb, C., "Die stacking (3D) microarchitecture", *39th Annual IEEE/ACM International Symposium on Microarchitecture*, Dec. 2006, pp. 469-479.

[15] Xu, T.; Liljeberg, P. & Tenhunen, H., "Exploring DRAM last level cache for 3D Network-on-Chip architecture", *Advanced Materials Research*, volume 403, 2010, pp. 4009-4018.

[16] Li, F.; Nicopoulos, C.; Richardson, T.; Xie, Y.; Narayanan, V. & Kandemir, M., "Design and management of 3D chip multiprocessors using network-in-memory", *Proceedings of the 33rd annual international symposium on Computer Architecture*, 2006, pp. 130-141.

[17] Jung, J.; Kang, K. & Kyung, C.-M., "Design and management of 3D-stacked NUCA cache for chip multiprocessors", *Proceedings of the 21st edition of the great lakes symposium on Great lakes symposium on VLSI*, May 2011, pp. 91-96.

[18] W. J. Dally and B. P. Towles, "Principles and Practices of Interconnection Networks", *Morgan Kaufmann*, 2004.

[19] J. L. Hennessy and D. A. Patterson, "Computer Architecture: A Quantitative Approach, 5th edition", *Morgan Kaufmann*, 2012.

[20] Binkert, N.; Beckmann, B.; Black, G.; Reinhardt, S. K.; Saidi, A.; Basu, A.; Hestness, J.; Hower, D. R.; Krishna, T.; Sardashti, S.; Sen, R.; Sewell, K.; Shoaib, M.; Vaish, N.; Hill, M. D. & Wood, D. A., "The gem5 simulator", *SIGARCH Computer Architecture News*, ACM, volume 39, issue 2, 2011, pp. 1-7.

[21] Bienia, C., "Benchmarking modern multiprocessors", Ph.D. Thesis, Princeton University, Jan. 2011.

[22] Kahng, A.; Li, B.; Peh, L.-S. & Samadi, K., "ORION 2.0: A fast and accurate NoC power and area model for early-stage design space exploration", *Design, Automation Test in Europe Conference Exhibition*, April 2009, pp. 423-428.