New Model Reduces Data Access Delay, Could increase Speed by up to 100x

Illinois Institute of Technology

CHICAGO, IL — As the amount of data grows ever larger but memory speed continues to greatly lag CPU speed, Xian-He Sun has established a new mathematical model for reducing data access delay. Sun is creator of Sun-Ni's law — one of three scalable computing laws along with Amdahl's law and Gustafson's law — and Distinguished Professor of Computer Science at Illinois Institute of Technology.

Called "Concurrent Average Memory Access Time (C-AMAT)," it promises to cut the penalty associated with accessing data and increase speed by up to 100 times through parallel memory access, which in turn will create a "break" in the memory-wall problem. A paper on C-AMAT is forthcoming in the Institute of Electrical and Electronics Engineers (IEEE) Computer Society's *Computer* magazine and can be viewed at http://www.computer.org/csdl/mags/co/preprint/06560068-abs.html.

"There's no question the primary limits on computing performance — from mobile phones to supercomputers — are the costs associated with data movement," said Andrew A. Chien, the William Eckhardt Professor in Computer Science and Senior Fellow in the Computation Institute at the University of Chicago, and Senior Computer Scientist at the Argonne National Laboratory. Before moving to University of Chicago, Chien was vice president of research at Intel.

"Dr. Sun's work attacks the critical problem of understanding and modeling data movement costs and systems performance and, thus, may enable better performing software (today) and improved hardware designs in the future."

During the last four decades, CPU speed increase has been following Moore's law, increasing 52 percent per year and doubling every 18 months. But memory speed is only increasing nine percent per year, and disk speed is even more behind, increasing an average of only six percent per year. Memory speed currently is about 400 times slower than CPU speed. That forms a wall for data movement and processing.

The data tsunami compounded with the memory-wall problem makes data management the primary concern of computing systems today, in terms of both performance and energy consumption. Computer science researchers increasingly are facing the need to rethink the design of computing systems from the conventional computing-centric view to a data-centric view.

While scientists search for solutions, they have found the value of the memory-bounded speedup model, or Sun-Ni's law, established by Xian-He Sun and Lionel Ni in 1990. In 1989, Intel developed the first processor with on-chip caches in response to the memory-wall problem. Cache is a small but fast special hardware device to hold data temporarily and make it more easily accessible.

In response, Sun and Ni introduced the memory-bounded formulation stating that the computing speed will be bounded by the on-chip memory or on-chip cache. A quantitative mathematical memory-bound function is presented for the tradeoff between memory, computing, and the effectiveness of the algorithm design in utilizing the cache architecture. Sun-Ni's law was introduced in advanced computer architecture textbooks in the 1990, and widely used in memory-bound-concerned algorithm design in the 2000s. In the 2010s, with the emergence of the big data problem, Sun-Ni's law appeared in popular magazines such as *PC Magazine.*

Sun's recent work on C-AMAT is the first formal mathematical model to promote and evaluate the concept of parallel memory for reducing data access delay via explicit parallel data access.

C-AMAT is a vital tool to mitigate the memory-wall effect and to improve memory system performance. There are only a handful fundamental formulas in computing architecture and algorithm design. AMAT is one of them. AMAT states that if the desired data is in cache (hit), then you get the data quickly; otherwise (miss), you get a cache miss penalty. Due to the memory-wall problem, the miss penalty will be big. So architecture and algorithm design focuses on reducing cache misses. With C-AMAT and parallel memory access, however, depending on if there is a hit occurring at the same time, a miss may or may not have a penalty. C-AMAT would change the focus of architecture and algorithm design from reducing cache misses to increasing data access parallelism. It provides a formulation to evaluate the effectiveness of the concurrency of each memory layer toward the final performance of parallel data access.

"The most profound research is not the design of the fastest algorithm for a given problem; it is revealing a fundamental computing property so hundreds or even thousands of algorithms can be developed upon it," Sun said.

Sun has been working on memory access issues for 20 years. During this period, his research has been continuously supported by the National Science Foundation (NSF) and other government agencies.

Sun's research is both application-driven and technology-driven. For the former, he and his research group have developed a series of software systems. These include the software packages of IOSIG: I/O Signatures Based Data Access Optimization, PFS-IOC: Server-side

I/O-Coordination in Parallel File System, GHS: Grid Harvest Service, and Network Bandwidth Predictor (NBP), etc. For the latter, his contributions include the abovementioned Sun-Ni Law (1990), C-AMAT (2013), the algorithm-machine combination scalability, the general speed-up metric, an extended Amdahl's law for multicore systems, and the memory Access Per Cycle (APC) performance metric, for measure memory parallelism (2011).

Sun is an IEEE Fellow for his contributions to memory-bounded performance metrics and scalable parallel computing, and a senior member of the Association for Computing Machinery (ACM). He has been the chair of the Department of Computer Science at IIT since fall 2009.