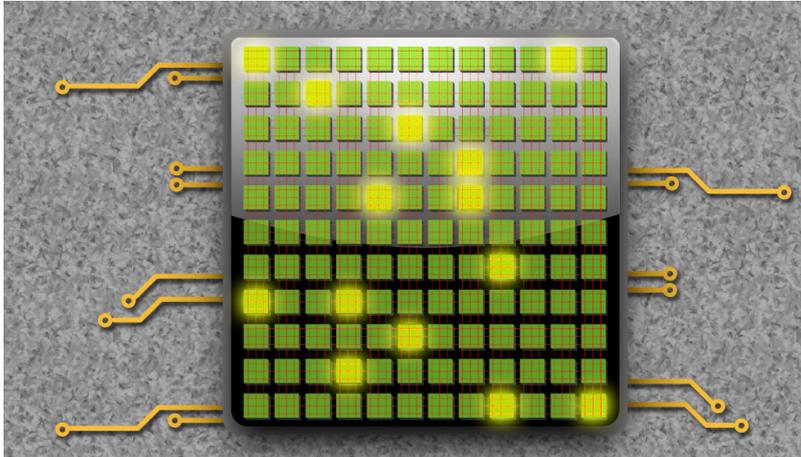


Clever Caches Improve Computer Chip Performance

MIT, Larry Hardesty



Computer chips keep getting faster because transistors keep getting smaller. But the chips themselves are as big as ever, so data moving around the chip, and between chips and main memory, has to travel just as far. As transistors get faster, the cost of moving data becomes, proportionally, a more severe limitation.

So far, chip designers have circumvented that limitation through the use of caches — small memory banks close to processors that store frequently used data. But the number of processors — or cores — per chip is also increasing, which makes cache management more difficult. Moreover, as cores proliferate, they have to share data more frequently, so the communication network connecting the cores becomes the site of more frequent logjams, as well.

In a pair of recent papers, researchers at [MIT](#) [1] and the [Univ. of Connecticut](#) [2] have developed a set of new caching strategies for massively multicore chips that, in simulations, significantly improved chip performance while actually reducing energy consumption.

The first paper, presented at the most recent ACM/IEEE International Symposium on Computer Architecture, reported average gains of 15 percent in execution time and energy savings of 25 percent. The second paper, which describes a complementary set of caching strategies and will be presented at the IEEE International Symposium on High Performance Computer Architecture, reports gains of 6 percent and 13 percent, respectively.

The caches on multicore chips are typically arranged in a hierarchy. Each core has its own private cache, which may itself have several levels, while all the cores share the so-called last-level cache, or LLC.

Chips' caching protocols usually adhere to the simple but surprisingly effective principle of "spatiotemporal locality." Temporal locality means that if a core

Clever Caches Improve Computer Chip Performance

Published on Laboratory Equipment (<http://www.laboratoryequipment.com>)

requests a particular piece of data, it will probably request it again. Spatial locality means that if a core requests a particular piece of data, it will probably request other data stored near it in main memory.

So every requested data item gets stored, along with those immediately adjacent to it, in the private cache. If it falls idle, it will eventually be squeezed out by more recently requested data, falling down through the hierarchy — from the private cache to the LLC to main memory — until it's requested again.

Different strokes

There are cases in which the principle of spatiotemporal locality breaks down, however. “An application works on a few, let's say, kilobytes or megabytes of data for a long period of time, and that's the working set,” says George Kurian, a graduate student in MIT's Department of Electrical Engineering and Computer Science and lead author on both papers. “One scenario where an application does not exhibit good spatiotemporal locality is where the working set exceeds the private-cache capacity.” In that case, Kurian explains, the chip could waste a lot of time cyclically swapping the same data between different levels of the cache hierarchy.

In the paper presented last year, Kurian; his advisor Srinivasa Devadas, the Edwin Sibley Webster Professor of Electrical Engineering and Computer Science at MIT; and Omer Khan, an assistant professor of electrical and computer engineering at the Univ. of Connecticut and a former postdoc in Devadas' lab, presented a hardware design that mitigates that problem. When an application's working set exceeds the private-cache capacity, the MIT researchers' chip would simply split it up between the private cache and the LLC. Data stored in either place would stay put, no matter how recently it's been requested, preventing a lot of fruitless swapping.

Conversely, if two cores working on the same data are constantly communicating in order to keep their cached copies consistent, the chip would store the shared data at a single location in the LLC. The cores would then take turns accessing the data, rather than clogging the network with updates.

The new paper examines the case where, to the contrary, two cores are working on the same data but communicating only infrequently. The LLC is usually treated as a single large memory bank: data stored in it is stored only once. But physically, it's distributed across the chip in discrete chunks. Kurian, Devadas and Khan have developed a second circuit that can treat these chunks, in effect, as extensions of the private cache. If two cores are working on the same data, each will receive its own copy in a nearby chunk of the LLC, enabling much faster data access.

Sentry box

The systems presented in both papers require active monitoring of the chips' operation — to determine, for instance, when working sets exceed some bound, or when multiple cores are accessing the same data. In each case, that monitoring requires a little extra circuitry, the equivalent of about 5 percent of the area of the

Clever Caches Improve Computer Chip Performance

Published on Laboratory Equipment (<http://www.laboratoryequipment.com>)

LLC. But, Kurian argues, because transistors keep shrinking, and communication isn't keeping up, chip space is not as crucial a concern as minimizing data transfer. Kurian, Devadas and Khan are also currently working to combine the two monitoring circuits, so that a single chip could deploy the cache-management strategies reported in both papers.

"It is a great piece of work," says Nikos Hardavellas, an assistant professor of electrical engineering and computer science at Northwestern Univ. "It definitely moves the state of the art forward." Existing caching schemes, Hardavellas explains, do treat different types of data differently: they might, for instance, use different caching strategies for program instructions and file data. "But if you dig deeper into these categories, you see that the data can behave very differently. In the past, we didn't know how to efficiently monitor the usefulness of the data. The [new] hardware design allows us to do this. That's a significant part of the contribution."

Moreover, Hardavellas says, "the two different designs seem to be working synergistically, which would indicate that the final result of combining the two would be better than the sum of the individual parts." As for commercialization of the technology, "I see no fundamental reason why not to," he says. "They seem implementable, they seem small enough, and they give us a significant benefit."

Source URL (retrieved on 05/31/2016 - 12:04pm):

<http://www.laboratoryequipment.com/news/2014/02/clever-caches-improve-computer-chip-performance>

Links:

[1] <http://web.mit.edu/>

[2] <http://uconn.edu/>