

Green Flash: Ultra-Efficient Climate Computing

More Science Using Less Power



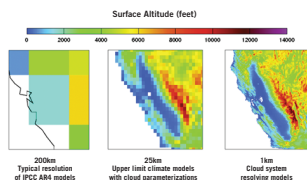
Berkeley Lab researchers have proposed an innovative way to improve global climate change predictions by using a supercomputer with low-power embedded microprocessors, an approach that would overcome limitations posed by today's conventional supercomputers.

A paper published in the May 2008 issue of the International Journal of High Performance Computing Applications lays out the benefit of a new class of supercomputers for modeling climate conditions and understanding climate change. Using the embedded microprocessor technology found in cell phones, iPods, toaster ovens, and many other modern-day electronic conveniences, they propose designing a cost-effective machine for running these models and improving climate predictions.

This research project has been named "Green Flash," after the optical phenomenon that sometimes appears on the horizon at sunset or sunrise.

Cloud-Resolving Climate Models Require Exascale Computing

The major source of error in current models is poor cloud simulation, and statistical cloud models are used to compensate for this. Accurately resolving clouds requires 1 km resolution — and this higher resolution also allows better comprehension of topography and violent storms. This transformational challenge is one of the driving applications for exascale computing.



Power Demand is the Major Challenge to Exascale

Processor	Clock	Peak/Core (GFlops)	Cores/Socket	Sockets	Cores	Power
AMD Opteron	2.8GHz	5.6	2	800K	1.7M	1180MW
IBM BGPP	850MHz	3.4	4	740K	3.0M	100MW
Green Flash/Tensilica Xtena	650MHz	2.7	32	120K	4.0M	51MW

The cost of power required to run HPC systems is projected to match or exceed hardware costs — Berkeley Lab's extrapolation of Blue Gene and AMD design trends shows energy demands of 100MW and 1180MW respectively, when applied to climate science.

"The only way to lower power consumption is to reduce waste."
—Mark Horowitz

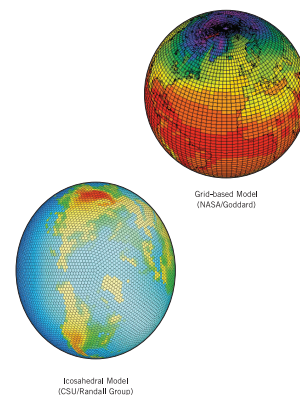
Application-Driven Architecture Design Improves Efficiency

Green Flash proposes tailoring the system architecture to climate modeling problems, replacing bloated serial processors with simple low-power cores. Borrowing ideas from embedded computing, the Green Flash team is co-designing hardware and software to gain greater efficiency. Using the Tensilica processor generator allows the fast creation of semi-customized cores, along with the compilers and other tools to use them.

New Algorithms are Needed for 1km Resolution Climate Model

The new model must run 1000 times faster than real time and must express 20 millionway parallelism. This cannot be achieved by running existing climate software and algorithms, which were designed for coarsegrained distributed memory architectures of the past.

The current grid-based model breaks down at 1km, which drives the need for an icosahedral model to reach 1km resolution.

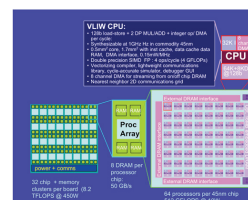
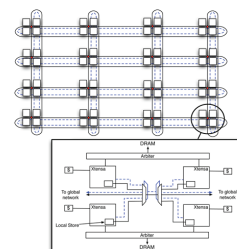


Low-Power Exascale Requires New Architectures

Achieving 1km resolution in climate models will require performance of 200 petaflop/s to 1 exaflop/s. Berkeley Lab's proposed architecture features 20 million processors, each at 500 Mflop/s sustained, combined with smart memories and an optical interconnect. The end result: 10 petaflop/s sustained, 100-200 petaflop/s peak performance.

- 100 terabytes total memory
- 5 GB/s local memory performance per domain (1 byte/flop)
- 200 MB/s in four nearest neighbor directions

Each physical processor maps to a sub-domain within the climate model, driving the need for new algorithms.



RAMP Provides Fast Hardware Emulation Using BEE3

The climate code is a long-running and highly complex workload. Current software performance simulators are slow and difficult to verify — and become intractable when high-flexibility, high-detail, and fast-execution are required! So, Green Flash is approaching the problem by using novel hardware emulation.

Green Flash leverages RAMP (the Research Accelerator for Multiple Processors), which is a multi-institutional effort to use reconfigurable logic for research into multi-processor architectures. RAMP has developed BEE3 (the Berkeley Emulation Engine, version 3), a 2U chassis with a tightly coupled 4 FPGA system for research computer architecture.

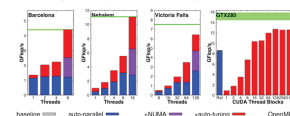
By using BEE3 to perform direct hardware emulation, the Green Flash prototype is fast enough to allow the execution of a full climate model in a short amount of time.



Auto Tuning Framework

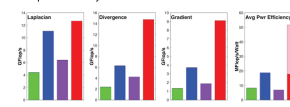
- Transitioning between different, complex architectures transparent to the user
- Allows performance portability across evolving architectural designs
- Domain specific knowledge allows for better optimization than traditional compilers can provide

Auto tuning increases performance, scalability ...



Effect of auto-tuning on performance of stencil kernels across multiple architectures. Note: The green marks the range in performance extrapolated from the Stream benchmark.

and power efficiency



Peak performance and power efficiency after auto-tuning. Note: GTC280 power efficiency is shown based on system power as well as the card alone.

K. Asanovic, D. Burke, D. Donofrio, L.A. Drummond, S. Kamil, C. McParland, N.L. Miller, M. Mohiyuddin, M. Murphy, L. Oliker, J. Shalf, J. Wawrzyniec, M.F. Wehner, K. Yelick

<http://www.lbl.gov/cs/html/greenflash.html>

