Parallel Computing & Accelerators

John Urbanic
Pittsburgh Supercomputing Center
Parallel Computing Scientist
Purpose of this talk

This is the 50,000 ft. view of the parallel computing landscape. We want to orient you a bit before parachuting you down into the trenches to deal with OpenACC. The plan is that you walk away with a knowledge of not just OpenACC, but also where it fits into the world of High Performance Computing.
### FLOPS we need: Climate change analysis

**Simulations**
- Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms
- New math, models, and systems support will be needed

**Extreme data**
- "Reanalysis" projects need 100× more computing to analyze observations
- Machine learning and other analytics are needed today for petabyte data sets
- Combined simulation/observation will empower policy makers and scientists

*Courtesy Horst Simon, LBNL*
Exascale combustion simulations

- Goal: 50% improvement in engine efficiency

- Center for Exascale Simulation of Combustion in Turbulence (ExaCT)
  - Combines M&S and experimentation
  - Uses new algorithms, programming models, and computer science

Courtesy Horst Simon, LBNL
Recent simulations achieve unprecedented scale of $65 \times 10^9$ neurons and $16 \times 10^{12}$ synapses.
The list is long, and growing.

- Molecular-scale Processes: atmospheric aerosol simulations
- AI-Enhanced Science: predicting disruptions in tokomak fusion reactors
- Hypersonic Flight
- Modeling Thermonuclear X-ray Bursts: 3D simulations of a neutron star surface or supernovae
- Quantum Materials Engineering: electrical conductivity photovoltaic and plasmonic devices
- Physics of Fundamental Particles: mass estimates of the bottom quark
- Digital Cells

These and others are in an appendix at the end of our Outro To Parallel Computing talk. And many of you doubtless brought your own immediate research concerns. Great!
Welcome to The Year of Exascale!

exa = $10^{18} = 1,000,000,000,000,000,000 = $\text{quintillion}$

64-bit precision floating point operations per second

Cray Red Storms 2004 (42 Tflops)
NVIDIA V100 (7.5 Tflops)
in very little time. Performing a billion operations, on the other hand, could take minutes or hours, though it’s still possible provided you are patient. Performing a trillion operations, however, will basically take forever. So a fair rule of thumb is that the calculations we can perform on a computer are ones that can be done with about a billion operations or less.
Where are those 10 or 12 orders of magnitude?

How do we get there from here?

BTW, that’s a bigger gap than
Moore's Law abandoned serial programming around 2004

Courtesy Liberty Computer Architecture Research Group
But Moore’s Law is only beginning to stumble now.

Intel process technology capabilities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Size</td>
<td>90nm</td>
<td>65nm</td>
<td>45nm</td>
<td>32nm</td>
<td>22nm</td>
<td>14nm</td>
<td>10nm</td>
<td>7nm</td>
</tr>
<tr>
<td>Integration Capacity</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
</tr>
</tbody>
</table>

Transistor for 90nm Process

Influenza Virus

Source: Intel

Source: CDC
But, at end of day we keep using getting more transistors.
That Power and Clock Inflection Point in 2004…
didn’t get better.

Fun fact: At 100+ Watts and <1V, currents are beginning to exceed 100A at the point of load!
Not a new problem, just a new scale...

Cray-2 with cooling tower in foreground, circa 1985
And how to get more performance from more transistors with the same power.

A 15% Reduction In Voltage Yields

**RULE OF THUMB**

<table>
<thead>
<tr>
<th>Frequency Reduction</th>
<th>Power Reduction</th>
<th>Performance Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>45%</td>
<td>10%</td>
</tr>
</tbody>
</table>

**SINGLE CORE**

- Area = 1
- Voltage = 1
- Freq = 1
- Power = 1
- Perf = 1

**DUAL CORE**

- Area = 2
- Voltage = 0.85
- Freq = 0.85
- Power = 1
- Perf = ~1.8

**Frequency Reduction**

15%

**Power Reduction**

45%

**Performance Reduction**

10%

A 15% Reduction In Voltage Yields

**SINGLE CORE**

- Area = 1
- Voltage = 1
- Freq = 1
- Power = 1
- Perf = 1

**DUAL CORE**

- Area = 2
- Voltage = 0.85
- Freq = 0.85
- Power = 1
- Perf = ~1.8
Parallel Computing

One woman can make a baby in 9 months.

Can 9 women make a baby in 1 month?

But 9 women can make 9 babies in 9 months.

First two bullets are Brook’s Law. From *The Mythical Man-Month*.

A must-read for serious project programmers that includes many other classics such as:

"What one programmer can do in one month, two programmers can do in two months."
Prototypical Application: Serial Weather Model
First Parallel Weather Modeling Algorithm: Richardson in 1917

Courtesy John Burkhardt, Virginia Tech
Weather Model: Shared Memory (OpenMP)

Fortran:

```fortran
!$omp parallel do
do i = 1, n
   a(i) = b(i) + c(i)
enddo
```

C/C++:

```c
#pragma omp parallel for
for(i=1; i<=n; i++)
a[i] = b[i] + c[i];
```
call MPI_Send( numbertosend, 1, MPI_INTEGER, index, 10, MPI_COMM_WORLD, errcode)

- 
- 

call MPI_Recv( numbertoreceive, 1, MPI_INTEGER, 0, 10, MPI_COMM_WORLD, status, errcode)

- 
- 
- 

call MPI_BARRIER(MPI_COMM_WORLD, errcode)

50 meteorologists using a telegraph.
Weather Model: Accelerator (OpenACC)

1 meteorologist coordinating 1000 math savants using tin cans and a string.
The pieces fit like this…

OpenMP

OpenACC

MPI
<table>
<thead>
<tr>
<th>#</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>CPU Interconnect [Accelerator]</th>
<th>Cores</th>
<th>Rmax (Tflops)</th>
<th>Rpeak (Tflops)</th>
<th>Power (MW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RIKEN Center for Computational Science Japan</td>
<td>Fujitsu</td>
<td>Fugaku</td>
<td>ARM 8.2A+ 48C 2.2GHz Torus Fusion Interconnect</td>
<td>7,299,072</td>
<td>442,010</td>
<td>537,212</td>
<td>29.8</td>
</tr>
<tr>
<td>2</td>
<td>DOE/SC/ORNL United States</td>
<td>IBM</td>
<td>Summit</td>
<td>Power9 22C 3.0 GHz Dual-rail Infiniband EDR NVIDIA V100</td>
<td>2,414,592</td>
<td>148,600</td>
<td>200,794</td>
<td>10.1</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LLNL United States</td>
<td>IBM</td>
<td>Sierra</td>
<td>Power9 3.1 GHz 22C Infiniband EDR NVIDIA V100</td>
<td>1,572,480</td>
<td>94,640</td>
<td>125,712</td>
<td>7.4</td>
</tr>
<tr>
<td>4</td>
<td>National Super Computer Center in Wuxi China</td>
<td>NRCP</td>
<td>Sunway TaihuLight</td>
<td>Sunway SW26010 260C 1.45GHz</td>
<td>10,649,600</td>
<td>93,014</td>
<td>125,435</td>
<td>15.3</td>
</tr>
<tr>
<td>5</td>
<td>DOE/LBNL/NERSC United States</td>
<td>HPE</td>
<td>Perlmutter</td>
<td>EPYC 64C 2.45 GHz Slingshot NVIDIA A100</td>
<td>761,304</td>
<td>70,870</td>
<td>93,750</td>
<td>2.6</td>
</tr>
<tr>
<td>6</td>
<td>NVIDIA Corp. United States</td>
<td>NVIDIA</td>
<td>Selene</td>
<td>EPYC 64C 2.25 GHz Infiniband HDR NVIDIA A100</td>
<td>555,520</td>
<td>63,460</td>
<td>79,215</td>
<td>2.6</td>
</tr>
<tr>
<td>7</td>
<td>National Super Computer Center in Guangzhou China</td>
<td>NUDT</td>
<td>Tianhe-2</td>
<td>Intel Xeon E5-2692 2.2 GHz TH Express-2 Intel Xeon Phi 31S1P</td>
<td>4,981,760</td>
<td>61,444</td>
<td>100,678</td>
<td>18.4</td>
</tr>
<tr>
<td>8</td>
<td>Forschungszentrum Juelich Germany</td>
<td>Bull</td>
<td>Juwels</td>
<td>EPYC 24C 2.8GHz Infiniband HDR NVIDIA A100</td>
<td>449,280</td>
<td>41,120</td>
<td>70,980</td>
<td>1.8</td>
</tr>
<tr>
<td>9</td>
<td>Eni S.p.A Italy</td>
<td>Dell</td>
<td>HPC5</td>
<td>Xeon 24C 2.1 GHz Infiniband HDR NVIDIA V100</td>
<td>669,760</td>
<td>35,450</td>
<td>51,720</td>
<td>2.2</td>
</tr>
<tr>
<td>10</td>
<td>Microsoft Azure East United States</td>
<td>MS Azure</td>
<td>Voyager</td>
<td>EPYC 48C 2.45 GHz InfiniBand HDR NVIDIA V100</td>
<td>253,440</td>
<td>30,050</td>
<td>39,531</td>
<td>1.8</td>
</tr>
</tbody>
</table>
The word is *Heterogeneous*

And it's not just supercomputers. It's on your desk, and in your phone.

How much of this can you program?
We can do better. We have a role model.

- Straightforward extrapolation results in a real-time human brain scale simulation at about 1 - 10 Exaflop/s with 4 PB of memory
- Exascale computers in 2022 will have a power consumption of at 20 - 30 MW
- The human brain takes 20W
- Even under best assumptions, our brain will still be a million times more power efficient

Courtesy Horst Simon, LBNL
Why you should be (extra) motivated.

- This parallel computing thing is no fad.
- The laws of physics are drawing this roadmap.
- If you get on board (the right bus), you can ride this trend for a long, exciting trip.

Let’s learn how to use these things!