“We used to joke that
“parallel computing is the future, and always will be,”
but the pessimists have been proven wrong.”

— Tony Hey
Course Information

- **Lecture Time:** MoFr 1:00 pm - 2:20 pm
- **Location:** Room 2120, Old CS Building, West Campus
- **Instructor:** Rezaul A. Chowdhury
- **Office Hours:** WeFr 4:00 pm - 5:30 pm, 239 New CS Building
- **Email:** rezaul@cs.stonybrook.edu
- **TA:** Unlikely

- **Class Webpage:**
Prerequisites

- **Required**: Background in algorithms analysis (e.g., CSE 373 or CSE 548)

- **Required**: Background in programming languages (C / C++)

- **Helpful but Not Required**: Background in computer architecture

- **Please Note**: This is not a course on
  - Programming languages
  - Computer architecture

- **Main Emphasis**: Parallel algorithms
Topics to be Covered

The following topics will be covered

– Analytical modeling of parallel programs
– Scheduling
– Programming using the message-passing paradigm and for shared address-space platforms
– Parallel algorithms for dense matrix operations, sorting, searching, graphs, computational geometry, and dynamic programming
– Concurrent data structures
– Transactional memory, etc.
Grading Policy

- Homeworks (three: lowest score 8%, highest score 20%, and the remaining one 12%): 40%

- Group project (one): 45%
  - Proposal: Feb 22
  - Progress report: Apr 1
  - Final demo/report: May 6 - May 10

- Scribe note (one lecture): 10%

- Class participation & attendance: 5%
Programming Environment

This course is supported by an educational grant from

- Extreme Science and Engineering Discovery Environment (XSEDE):
  https://www.xsede.org

We have access to the following supercomputing resources

- **Stampede 2 (Texas Advanced Computing Center):**
  4,200 KNL nodes with 68 cores (Intel Xeon Phi 7250 / “Knights Landing”) each;
  1,736 SKX nodes each with 48 cores (Intel Xeon Platinum / “Skylake”) on two sockets.

- **Comet (San Diego Supercomputer Center):**
  1,984 nodes with 24 cores (2 Intel Haswell) per node. The Comet GPU resource features 36 K80 GPU nodes (with 2 Intel Haswell processors each), and 36 P100 nodes (with 2 Intel Broadwell processors each).
# World’s Most Powerful Supercomputers in November, 2018

(www.top500.org)

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax [TFlop/s]</th>
<th>Rpeak [TFlop/s]</th>
<th>Power [kW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States</td>
<td>2,397,824</td>
<td>143,500.0</td>
<td>200,794.9</td>
<td>9,783</td>
</tr>
<tr>
<td>2</td>
<td>Sierra - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNl United States</td>
<td>1,572,480</td>
<td>96,460.0</td>
<td>125,712.0</td>
<td>7,438.3</td>
</tr>
<tr>
<td>3</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRPC National Supercomputing Center in Wuxi China</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,635.9</td>
<td>15,371</td>
</tr>
<tr>
<td>4</td>
<td>Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUJT National Super Computer Center in Guangzhou China</td>
<td>4,981,760</td>
<td>61,444.5</td>
<td>100,678.7</td>
<td>18,482</td>
</tr>
<tr>
<td>5</td>
<td>Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc. Swiss National Supercomputing Centre [CSCS] Switzerland</td>
<td>387,872</td>
<td>21,230.0</td>
<td>27,154.3</td>
<td>2,364.2</td>
</tr>
<tr>
<td>6</td>
<td>Trinity - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect, Cray Inc. DOE/NNSA/LANL/SNL United States</td>
<td>979,072</td>
<td>20,158.7</td>
<td>41,461.2</td>
<td>7,578.1</td>
</tr>
<tr>
<td>7</td>
<td>Al Bridging Cloud Infrastructure (ABCI) - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology [AIST] Japan</td>
<td>391,680</td>
<td>19,880.0</td>
<td>32,576.6</td>
<td>1,649.3</td>
</tr>
<tr>
<td>8</td>
<td>SuperMUC-NG - ThinkSystem SD530, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path, Lenovo Leibniz Rechenzentrum Germany</td>
<td>305,856</td>
<td>19,476.6</td>
<td>26,873.9</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Titan - Cray XK7, Opteron 6274 16C 2.20GHz, Cray Gemini interconnect, NVIDIA K20x, Cray Inc. DOE/SC/Oak Ridge National Laboratory United States</td>
<td>568,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,209</td>
</tr>
<tr>
<td>10</td>
<td>Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States</td>
<td>1,572,864</td>
<td>17,173.2</td>
<td>20,132.7</td>
<td>7,890</td>
</tr>
</tbody>
</table>
## World’s Most Powerful Supercomputers in November, 2018

(www.top500.org)

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Lassen - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100, IBM / NVIDIA / Mellanox DOE/NNSA/LNL</td>
<td>248,976</td>
<td>15,430.0</td>
<td>19,904.4</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Cori – Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/SC/LBNL/NERSC</td>
<td>622,336</td>
<td>14,014.7</td>
<td>27,880.7</td>
<td>3,939</td>
</tr>
<tr>
<td>13</td>
<td>Nuniform - Cray CS500, Intel XeonPhi 7250 68C 1.4GHz, Intel Omni-Path , Cray Inc. Korea Institute of Science and Technology Information Korea, South</td>
<td>570,020</td>
<td>13,929.3</td>
<td>25,705.9</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path, Fujitsu Joint Center for Advanced High Performance Computing Japan</td>
<td>556,104</td>
<td>13,554.6</td>
<td>24,913.5</td>
<td>2,718.7</td>
</tr>
<tr>
<td>15</td>
<td>HPCC - ProLiant DL380 Gen10, Xeon Platinum 8160 24C 2.1GHz, Mellanox InfiniBand EDR, NVIDIA Tesla P100, HPE Eni S.p.A Italy</td>
<td>253,600</td>
<td>12,210.0</td>
<td>18,621.1</td>
<td>1,320</td>
</tr>
<tr>
<td>16</td>
<td>Tera-1000-2 - Bull Sequana X1000, Intel Xeon Phi 7250 68C 1.4GHz, Bull BX1 1.2 , Bull, Atos Group Commissariat a l’Energie Atomique [CEA] France</td>
<td>561,408</td>
<td>11,965.5</td>
<td>23,396.4</td>
<td>3,178</td>
</tr>
<tr>
<td>17</td>
<td>Stampede2 - PowerEdge C6320P/C6420, Intel Xeon Phi 7250 68C 1.4GHz/Platinum 8160, Intel Omni-Path , Dell EMC Texas Advanced Computing Center/Univ. of Texas United States</td>
<td>367,024</td>
<td>10,680.7</td>
<td>18,309.2</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>K computer; SPARC64 VIIIfx 2.0GHz, Tofu interconnect , Fujitsu RIKEN Advanced Institute for Computational Science [AICS] Japan</td>
<td>705,024</td>
<td>10,510.0</td>
<td>11,280.4</td>
<td>12,659.9</td>
</tr>
<tr>
<td>19</td>
<td>Marconi Intel Xeon Phi - CINECA Cluster, Lenovo SDS30/S720AP, Intel Xeon Phi 7250 68C 1.4GHz/Platinum 8160, Intel Omni-Path , Lenovo CINECA Italy</td>
<td>348,000</td>
<td>10,384.9</td>
<td>10,816.0</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Taiwania 2 - QCT QuantaGrid D52G-4U/LC, Xeon Gold 6154 18C 3GHz, Mellanox InfiniBand EDR, NVIDIA Tesla V100 SXM2 , Quanta Computer / Taiwan Fixed Network / ASUS Cloud National Center for High Performance Computing Taiwan</td>
<td>170,352</td>
<td>9,600.0</td>
<td>15,208.2</td>
<td>797.5</td>
</tr>
</tbody>
</table>
## Programming Environment

### World’s Most Powerful Supercomputers in November, 2017

( [www.top500.org](http://www.top500.org) )

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi, China</td>
<td>10,649,600</td>
<td>93,914.6</td>
<td>125,435.9</td>
<td>15,371</td>
</tr>
<tr>
<td>2</td>
<td>Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.20GHz, TH Express-2, Intel Xeon Phi 3151P, NUDT National Super Computer Center in Guangzhou, China</td>
<td>3,120,000</td>
<td>33,862.7</td>
<td>54,902.4</td>
<td>17,908</td>
</tr>
<tr>
<td>3</td>
<td>Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.60GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc, Swiss National Supercomputing Centre (CSCS) Switzerland</td>
<td>361,740</td>
<td>19,590.0</td>
<td>25,326.3</td>
<td>2,272.0</td>
</tr>
<tr>
<td>4</td>
<td>Gyoukou - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.30GHz, InfiniBand EDR, PEZY-SC2 700MHz, ExaScaler Japan Agency for Marine Earth Science and Technology Japan</td>
<td>19,860,000</td>
<td>19,135.8</td>
<td>28,192.0</td>
<td>1,350.2</td>
</tr>
<tr>
<td>5</td>
<td>Titan - Cray XK7, Opteron 6274 16C 2.20GHz, Cray Gemini interconnect, NVIDIA K20x, Cray Inc. DOE/SC/Oak Ridge National Laboratory United States</td>
<td>560,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,209</td>
</tr>
<tr>
<td>6</td>
<td>Sequoia - BlueGene/Q, Power BDC 16C 1.60 GHz, Custom, IBM DOE/NNSA/LINL United States</td>
<td>1,572,864</td>
<td>17,173.2</td>
<td>20,132.7</td>
<td>7,890</td>
</tr>
<tr>
<td>7</td>
<td>Trinity - Cray XC40, Intel Xeon Phi 7250 68C 1.40GHz, Aries interconnect, Cray Inc. DOE/NNSA/ANL/JSNL United States</td>
<td>979,916</td>
<td>14,137.3</td>
<td>43,902.6</td>
<td>3,843.6</td>
</tr>
<tr>
<td>8</td>
<td>Cori - Cray XC40, Intel Xeon Phi 7250 68C 1.40GHz, Aries interconnect, Cray Inc. DOE/SC/NL/NERSC United States</td>
<td>622,336</td>
<td>14,014.7</td>
<td>27,800.7</td>
<td>3,939</td>
</tr>
<tr>
<td>9</td>
<td>Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.40GHz, Intel Omni-Path, Fujitsu Joint Center for Advanced High Performance Computing Japan</td>
<td>556,104</td>
<td>13,554.6</td>
<td>24,913.5</td>
<td>2,718.7</td>
</tr>
<tr>
<td>10</td>
<td>K computer, SPARC64 VIIIfx 2.00GHz, Tofu interconnect, Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan</td>
<td>705,024</td>
<td>10,510.0</td>
<td>11,280.4</td>
<td>12,659.9</td>
</tr>
<tr>
<td>11</td>
<td>Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom, IBM DOE/SC/Norwegian National Laboratory United States</td>
<td>786,432</td>
<td>8,586.6</td>
<td>10,066.3</td>
<td>3,945</td>
</tr>
<tr>
<td>12</td>
<td>Stampede-2 - PowerEdge C6320PG6420, Intel Xeon Phi 7250 68C 1.44GHz/Platinum 8160, Intel Omni-Path, Dell EMC Texas Advanced Computing Center/Univ. of Texas United States</td>
<td>368,928</td>
<td>8,317.7</td>
<td>18,215.8</td>
<td>8,445</td>
</tr>
</tbody>
</table>
Recommended Textbooks


Why Parallelism?
Unicore Performance Has Hit a Wall!

Some Reasons

- Lack of additional ILP
  (Instruction Level Hidden Parallelism)
- High power density
- Manufacturing issues
- Physical limits
- Memory speed
Unicore Performance: No Additional ILP

“Everything that can be invented has been invented.”

— Charles H. Duell
Commissioner, U.S. patent office, 1899

Exhausted all ideas to exploit hidden parallelism?

– Multiple simultaneous instructions
– Instruction Pipelining
– Out-of-order instructions
– Speculative execution
– Branch prediction
– Register renaming, etc.
ILP: Instruction Pipelining

**Unicore Performance: High Power Density**

- Dynamic power, $P_d \propto V^2 f C$
  - $V = $ supply voltage
  - $f = $ clock frequency
  - $C = $ capacitance

- But $V \propto f$
- Thus $P_d \propto f^3$

*Source: Patrick Gelsinger, Intel Developer Forum, Spring 2004 (Simon Floyd)*
Unicore Performance: Manufacturing Issues

— Frequency, \( f \propto \frac{1}{s} \)
  
  — \( s = \text{feature size (transistor dimension)} \)

— Transistors / unit area \( \propto \frac{1}{s^2} \)

— Typically, die size \( \propto \frac{1}{s} \)

— So, what happens if feature size goes down by a factor of \( x \)?
  
  — Raw computing power goes up by a factor of \( x^4 \)!
  
  — Typically most programs run faster by a factor of \( x^3 \)
    without any change!

Source: Kathy Yelick and Jim Demmel, UC Berkeley
Unicore Performance: Manufacturing Issues

— Manufacturing cost goes up as feature size decreases
  — Cost of a semiconductor fabrication plant doubles every 4 years (Rock’s Law)
— CMOS feature size is limited to 5 nm (at least 10 atoms)

Source: Kathy Yelick and Jim Demmel, UC Berkeley
Unicore Performance: Physical Limits

Execute the following loop on a serial machine in 1 second:

```c
for ( i = 0; i < 10^{12}; ++i )
    z[ i ] = x[ i ] + y[ i ];
```

- We will have to access $3 \times 10^{12}$ data items in one second
- Speed of light is, $c \approx 3 \times 10^8$ m/s
- So each data item must be within $c / 3 \times 10^{12} \approx 0.1$ mm from the CPU on the average
- All data must be put inside a $0.2 \text{ mm} \times 0.2 \text{ mm}$ square
- Each data item ($\geq 8$ bytes) can occupy only 1 Å$^2$ space! (size of a small atom!)

*Source:* Kathy Yelick and Jim Demmel, UC Berkeley
Unicore Performance: Memory Wall

Source: Rick Hetherington, Chief Technology Officer, Microelectronics, Sun Microsystems
Unicore Performance Has Hit a Wall!

Some Reasons

– Lack of additional ILP
  ( Instruction Level Hidden Parallelism )

– High power density

– Manufacturing issues

– Physical limits

– Memory speed

“Oh Sinnerman, where you gonna run to?”

― Sinnerman (recorded by Nina Simone)
Where You Gonna Run To?

– Changing $f$ by 20% changes performance by 13%

– So what happens if we overclock by 20%?

Source: Andrew A. Chien, Vice President of Research, Intel Corporation
Where You Gonna Run To?

– Changing $f$ by 20% changes performance by 13%
– So what happens if we overclock by 20%?
– And underclock by 20%?

Source: Andrew A. Chien, Vice President of Research, Intel Corporation
Changing $f$ by 20% changes performance by 13%

So what happens if we overclock by 20%?

And underclock by 20%?

Source: Andrew A. Chien, Vice President of Research, Intel Corporation
Moore’s Law Reinterpreted

Source: Report of the 2011 Workshop on Exascale Programming Challenges
Top 500 Supercomputing Sites (Cores / Socket)

www.top500.org
(Nov, 2017)
No Free Lunch for Traditional Software

Source: Simon Floyd, Workstation Performance: Tomorrow's Possibilities (Viewpoint Column)
Insatiable Demand for Performance

Weather Prediction
Oil Exploration
Design Simulation
Genomics Research
Financial Analysis
Medical Imaging

Some Useful Classifications of Parallel Computers
Parallel Computer Memory Architecture (Distributed Memory)

- Each processor has its own local memory — no global address space
- Changes in local memory by one processor have no effect on memory of other processors
- Communication network to connect inter-processor memory
- Programming
  - Message Passing Interface (MPI)
  - Many once available: PVM, Chameleon, MPL, NX, etc.

Source: Blaise Barney, LLNL
Parallel Computer Memory Architecture (Shared Memory)

- All processors access all memory as global address space
- Changes in memory by one processor are visible to all others
- Two types
  - Uniform Memory Access (UMA)
  - Non-Uniform Memory Access (NUMA)
- Programming
  - Open Multi-Processing (OpenMP)
  - Cilk/Cilk++ and Intel Cilk Plus
  - Intel Thread Building Block (TBB), etc.

Source: Blaise Barney, LLNL
Parallel Computer Memory Architecture (Hybrid Distributed-Shared Memory)

- The share-memory component can be a cache-coherent SMP or a Graphics Processing Unit (GPU)
- The distributed-memory component is the networking of multiple SMP/GPU machines
- Most common architecture for the largest and fastest computers in the world today
- Programming
  - OpenMP / Cilk + CUDA / OpenCL + MPI, etc.

Source: Blaise Barney, LLNL
Flynn’s Taxonomy of Parallel Computers

Flynn’s classical taxonomy (1966):

Classification of multi-processor computer architectures along two independent dimensions of instruction and data.

<table>
<thead>
<tr>
<th>Single Instruction (SI)</th>
<th>Single Data (SD)</th>
<th>Multiple Data (MD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SISD</td>
<td>SIMD</td>
</tr>
<tr>
<td>Multiple Instruction (MI)</td>
<td>MISD</td>
<td>MIMD</td>
</tr>
</tbody>
</table>
Flynn’s Taxonomy of Parallel Computers

SISD

- A serial (non-parallel) computer
- The oldest and the most common type of computers
- Example: Uniprocessor unicore machines

Source: Blaise Barney, LLNL
Flynn’s Taxonomy of Parallel Computers

SIMD

- A type of parallel computer
- All PU’s run the same instruction at any given clock cycle
- Each PU can act on a different data item
- Synchronous (lockstep) execution
- Two types: processor arrays and vector pipelines
- Example: GPUs (Graphics Processing Units)

Source: Blaise Barney, LLNL
Flynn’s Taxonomy of Parallel Computers

MISD
- A type of parallel computer
- Very few ever existed

MIMD
- A type of parallel computer
- Synchronous /asynchronous execution
- Examples: most modern supercomputers, parallel computing clusters, multicore PCs

Source: Blaise Barney, LLNL
“The way the processor industry is going, is to add more and more cores, but nobody knows how to program those things. I mean, two, yeah; four, not really; eight, forget it.”

— Steve Jobs, NY Times interview, June 10 2008
Consider the following loop:

\[ \text{for } i = 1 \text{ to } n \text{ do } \]
\[ C[i] \leftarrow A[i] \times B[i] \]

- Suppose you have an infinite number of processors/cores
- Ignore all overheads due to scheduling, memory accesses, communication, etc.
- Suppose each operation takes a constant amount of time
- How long will this loop take to complete execution?
Consider the following loop:

\[
\text{for } i = 1 \text{ to } n \text{ do } \\
C[i] \leftarrow A[i] \times B[i]
\]

- Suppose you have an infinite number of processors/cores
- Ignore all overheads due to scheduling, memory accesses, communication, etc.
- Suppose each operation takes a constant amount of time
- How long will this loop take to complete execution?
  - $O(1)$ time
Now consider the following loop:

\[
c \leftarrow 0
\]

\[
\text{for } i = 1 \text{ to } n \text{ do}
\]

\[
c \leftarrow c + A[i] \times B[i]
\]

— How long will this loop take to complete execution?
Parallel Algorithms Warm-up (2)

Now consider the following loop:

\[
\begin{aligned}
c &\leftarrow 0 \\
for i = 1 \text{ to } n \text{ do } \\
c &\leftarrow c + A[i] \times B[i]
\end{aligned}
\]

— How long will this loop take to complete execution?

— \(O(\log n)\) time
Parallel Algorithms Warm-up (3)

Now consider quicksort:

\[
QSort( A )
\]

\[
if \ |A| \leq 1 \ return \ A
\]

\[
else \ p \leftarrow A[ \ rand( |A| ) ]
\]

\[
return \ QSort( \ \{ x \in A: x < p \} )
\]

\[
\# \ \{ p \} \ #
\]

\[
QSort( \ \{ x \in A: x > p \} )
\]

— Assuming that \( A \) is split in the middle everytime, and the two recursive calls can be made in parallel, how long will this algorithm take?
Now consider quicksort:

\[
QSort( A ) \\
\text{if } |A| \leq 1 \text{ return } A \\
\text{else } p \leftarrow A[ \text{rand}( |A| ) ] \\
\text{return } QSort( \{ x \in A : x < p \} ) \\
\quad \# \{ p \} \# \\
\quad QSort( \{ x \in A : x > p \} )
\]

— Assuming that \( A \) is split in the middle every time, and the two recursive calls can be made in parallel, how long will this algorithm take?

— \( O( \log^2 n ) \) (if partitioning takes logarithmic time)
— \( O( \log n ) \) (but can be partitioned in constant time)