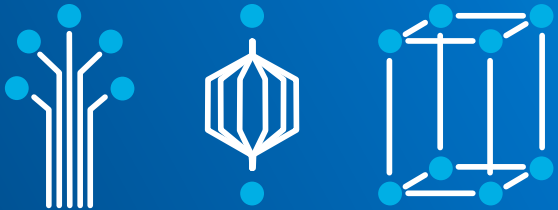
Fundamentals of Parallelism on Intel® Architecture

Week 1
Modern Code



COLFAX RESEARCH

May 2017

intel

Course Roadmap

Lecture 1 Intel Architecture and Modern Code

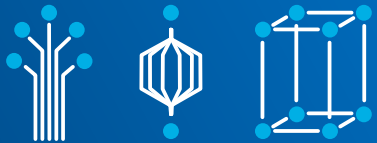
Lecture 2 Multi-Threading with OpenMP

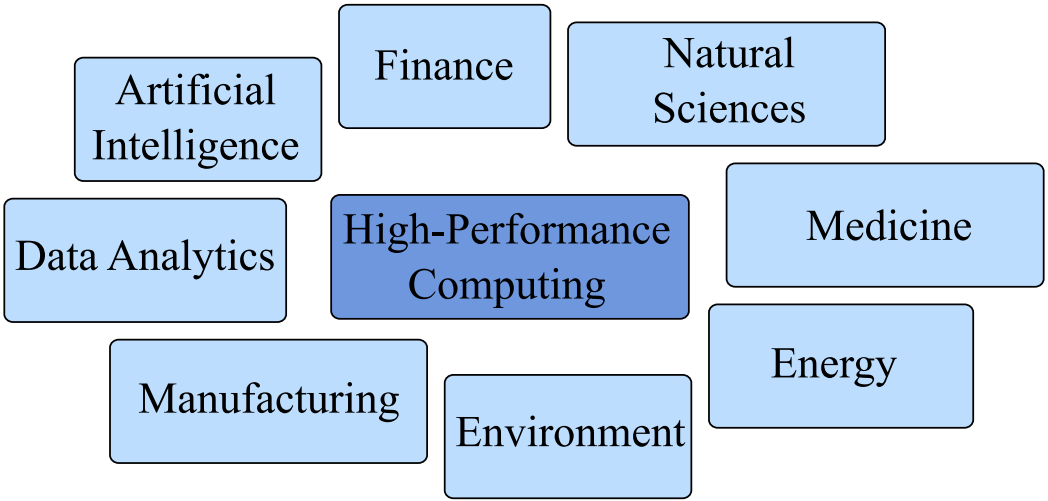
Lecture 3 Vectorization with Intel Compilers

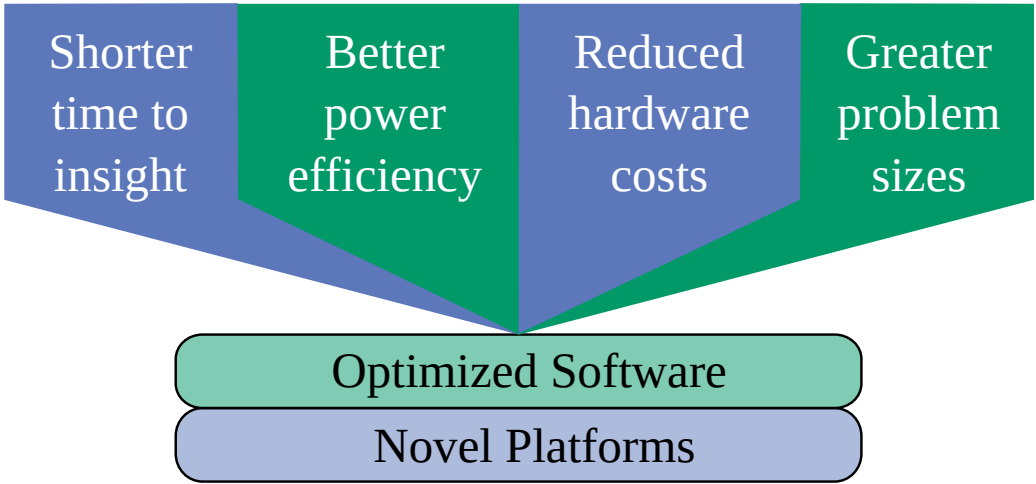
Lecture 4 Essentials of Memory Traffic

Lecture 5 Clusters and MPI

§1. Why This Course

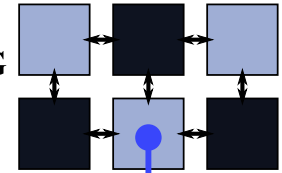



Artificial Intelligence
Finance
Natural Sciences
Data Analytics
High-Performance Computing
Medicine
Manufacturing
Environment
Energy


Shorter time to insight
Better power efficiency
Reduced hardware costs
Greater problem sizes
Optimized Software
Novel Platforms

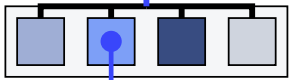
Parallel Programming Layers

CLUSTER COMPUTING
in distributed memory



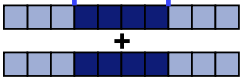
```
MPI_Sendrecv(data, k,
   MPI_DOUBLE, data2,
   ... );
```

MULTITHREADING
in shared memory



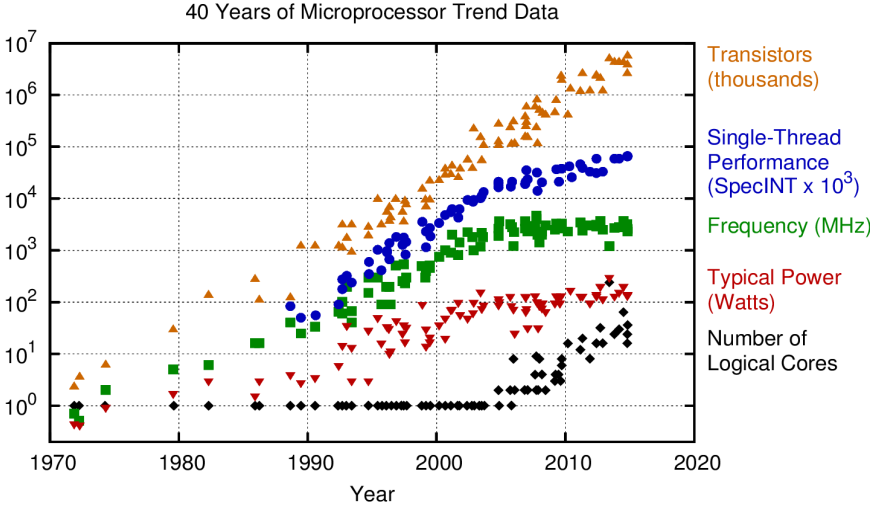
```
#pragma omp parallel for
for (j = 0; j < m; j++)
  ComputeSubset(j);
```

VECTORIZATION
of floating-point math



```
#pragma omp simd
for (i = 0; i < n; i++)
  A[i] += B[i];
```

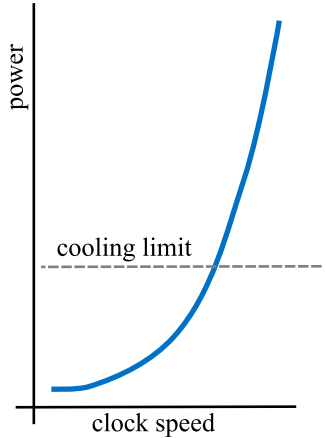
§2. How Computers Get Faster



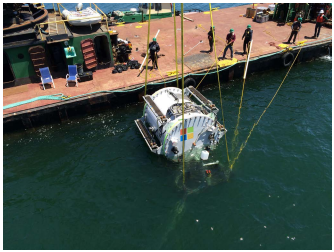
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Source: karlrupp.net

Clock Speed and Power Wall



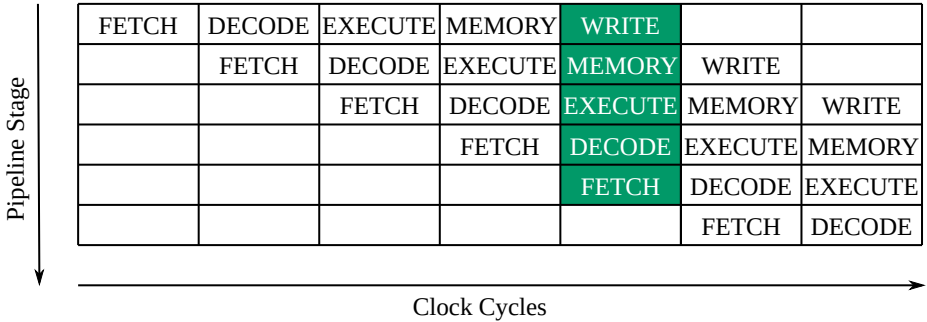
Overclocking?



Free cooling?

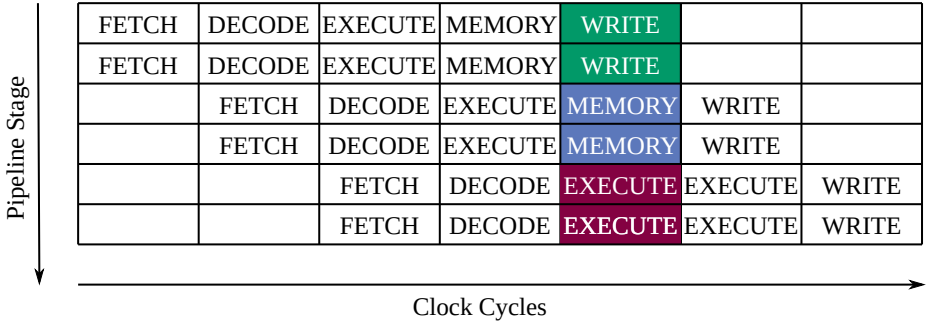
Cooling solutions for high clock speeds are not practical or expensive

Pipelining and ILP Wall



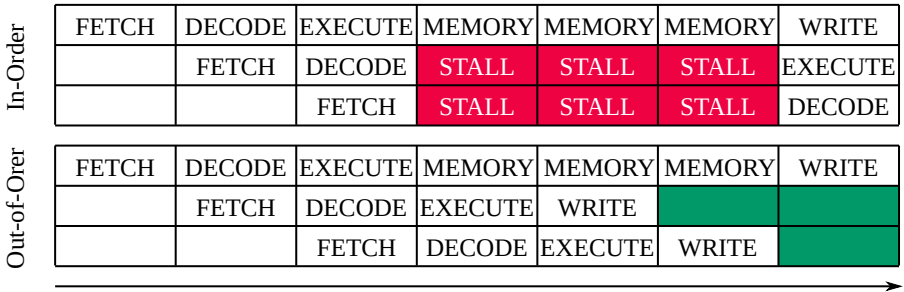
Only so many pipeline stages, possible conflicts

Superscalar Execution and ILP Wall



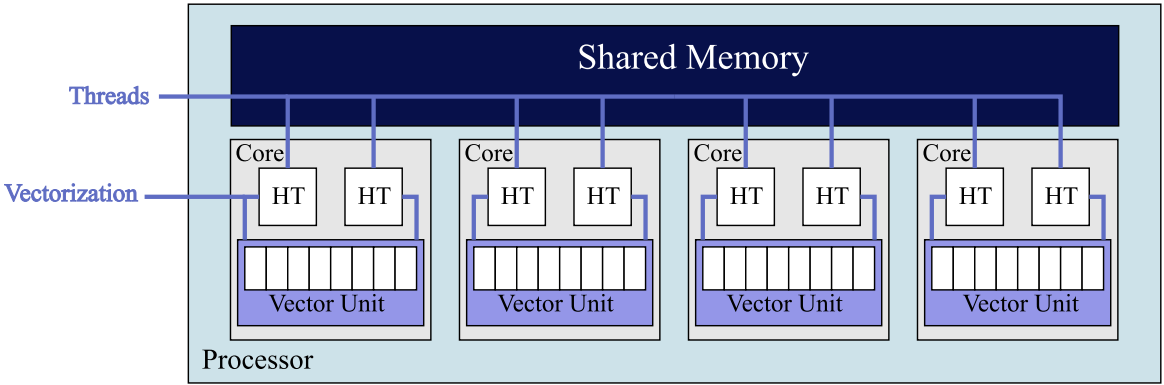
Automatic search for independent instructions reqiures extra resources

Out-of-Order Execution and Memory Wall



Limited by data locality access pattern in application.

Parallelism: Cores and Vectors



Unbounded growth opportunity, but **not automatic**

Parallelism is the Path Forward

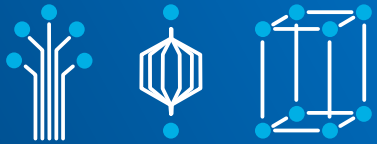
- Power wall
- ILP wall
- Memory wall

The Show Must Go On

Hardware keeps evolving through parallelism.

Software must catch up!

§3. Intel Architecture



Intel Computing Platforms

General-Purpose Processors

Intel® Xeon®
Intel® Core™
Intel® Atom™, ...

Specialized Processors

Intel® Xeon Phi™ processors and coprocessors

Computing Accelerators

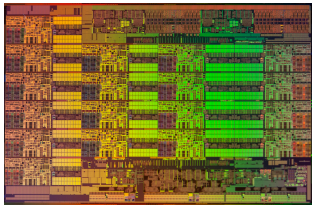
Intel® VCA (x86)
Intel® Nervana™ Platform
Intel® DLIA™ (FPGAs)

Network Interconnects

Intel® Omni-Path™ Architecture

Intel Xeon Processors

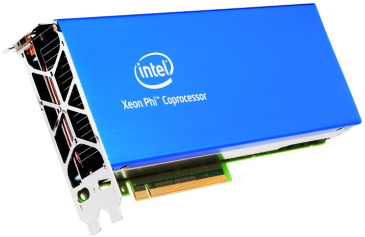
- 1-, 2-, 4-way
- General-purpose
- Highly parallel (44 cores*)
- Resource-rich
- Forgiving performance
- Theor. ~ 1.0 TFLOP/s in DP*
- Meas. ~ 154 GB/s bandwidth*

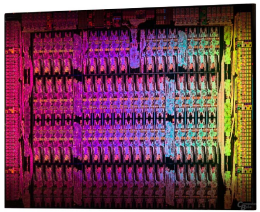


* 2-way Intel Xeon processor, Broadwell architecture (2016), top-of-the-line (e.g., E5-2699 V4)

Intel Xeon Phi Processors (1st Gen)

- PCIe add-in card
- Specialized for computing
- Highly-parallel (61 cores*)
- Balanced for compute
- Less forgiving
- Theor. ~ 1.2 TFLOP/s in DP*
- Meas. ~ 176 GB/s bandwidth*

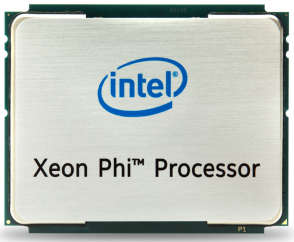


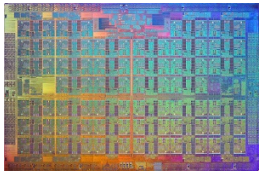


* Intel Xeon Phi coprocessor, Knighs Corner architecture (2012), top-of-the-line (e.g., 7120P)

Intel Xeon Phi Processors (2nd Gen)

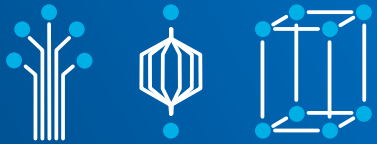
- Bootable or PCIe add-in card
- Specialized for computing
- Highly-parallel (72 cores*)
- Balanced for compute
- Less forgiving than Xeon
- Theor. ~ 3.0 TFLOP/s in DP*
- Meas. ~ 490 GB/s bandwidth*



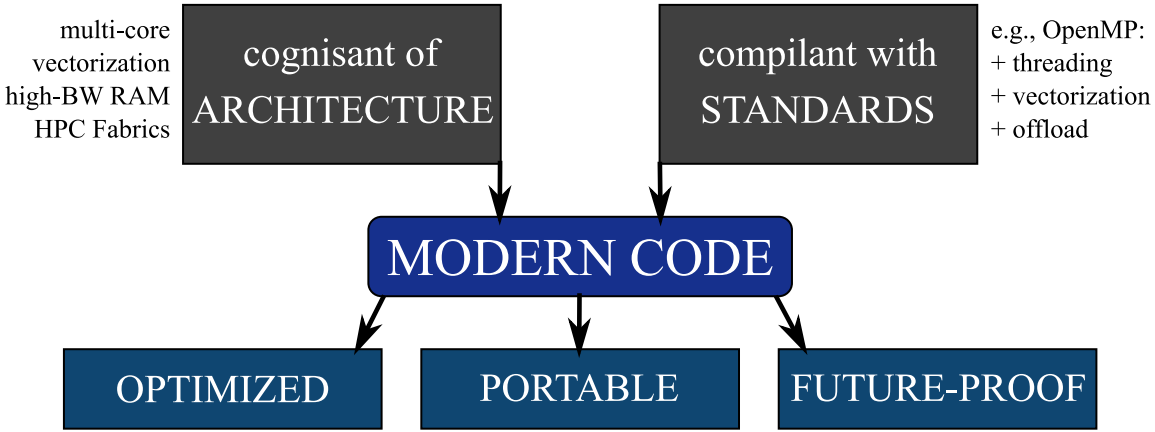


* Intel Xeon Phi processor, Knighs Landing architecture (2016), top-of-the-line (e.g., 7290P)

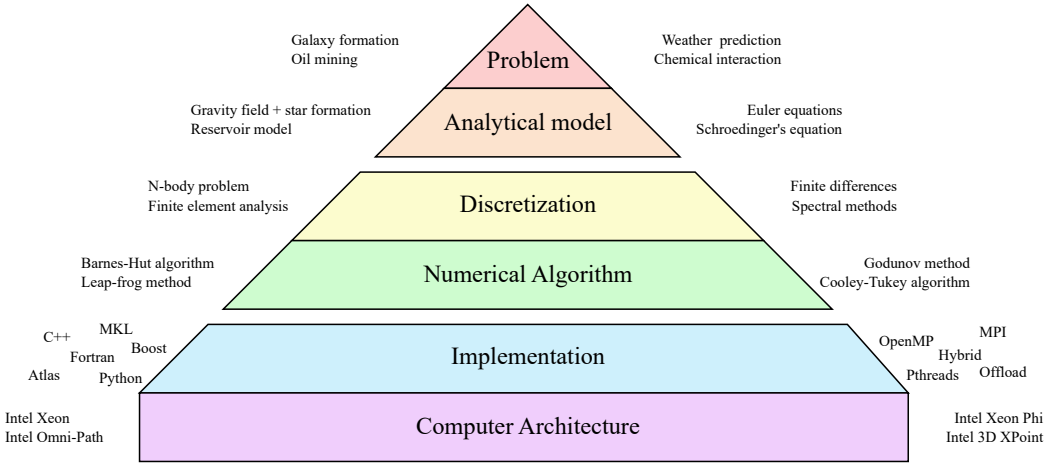
§4. Modern Code



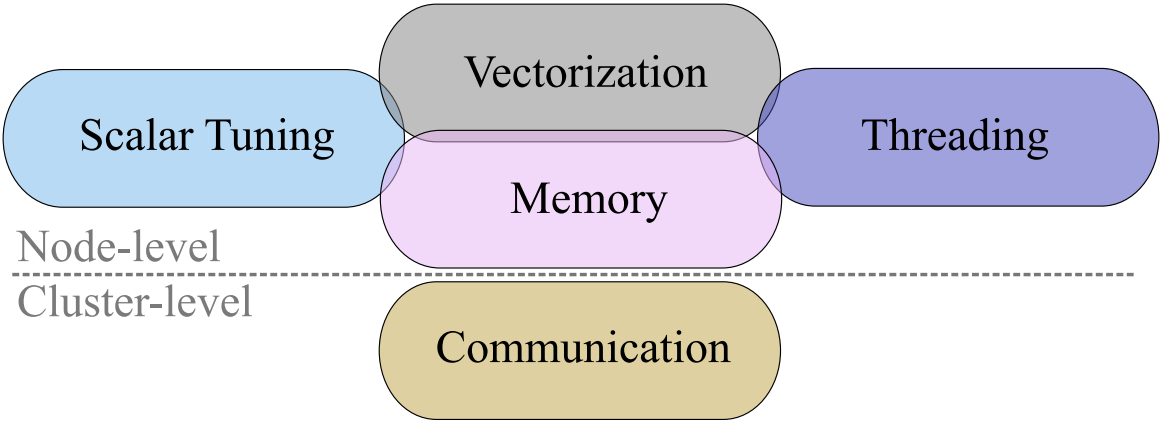
One Code for All Platforms



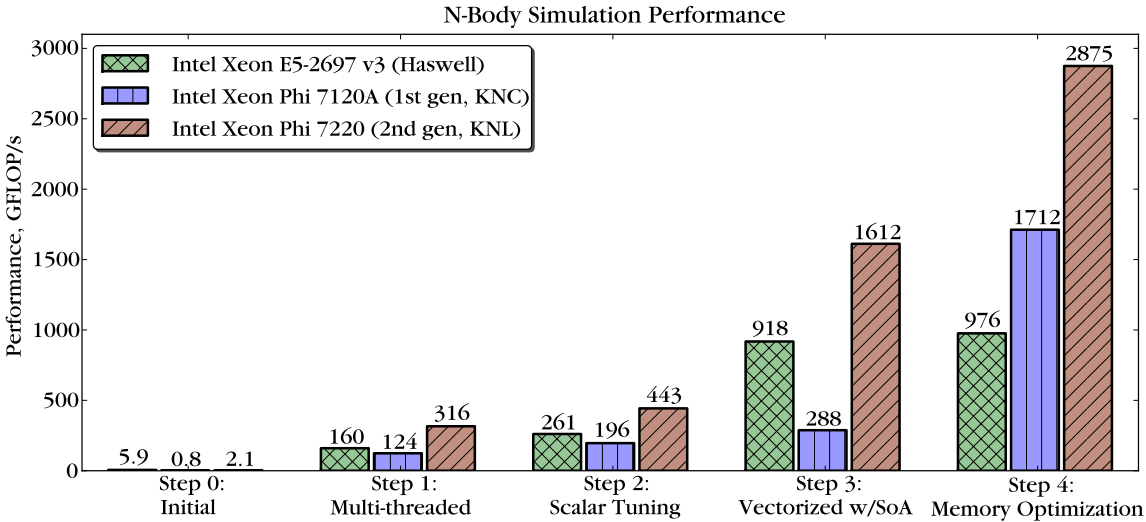
Computing in Science and Engineering



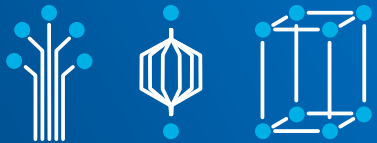
Optimization Areas



Common Code Modernization Experience

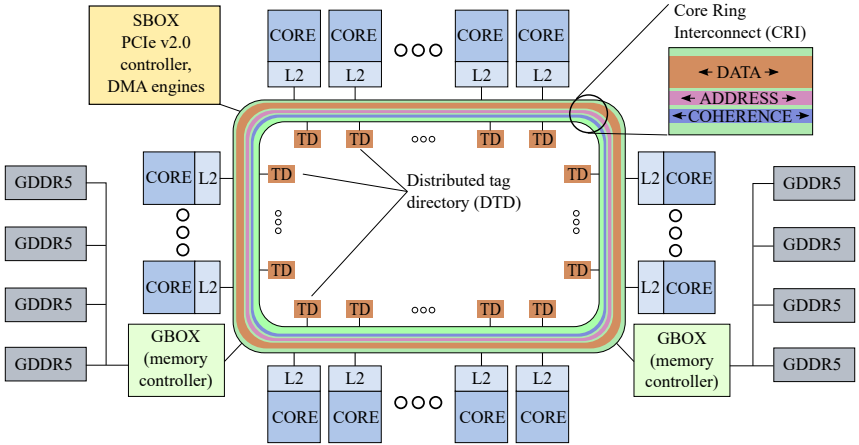


§5. What You Are Going to Learn



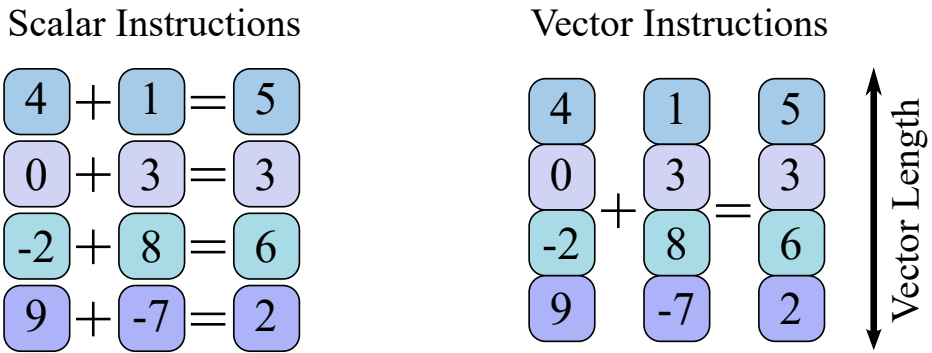
Lecture 2: Multi-Threading with OpenMP

Cores implement MIMD (Multiple Instruction Multiple Data) arch



Lecture 3: Vectorization with Intel Compilers

Vectors – form Single Instruction Multiple Data (SIMD) architecture

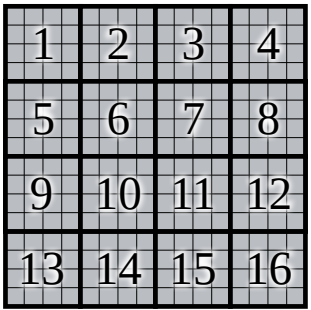


Lecture 4: Essentials of Memory Traffic

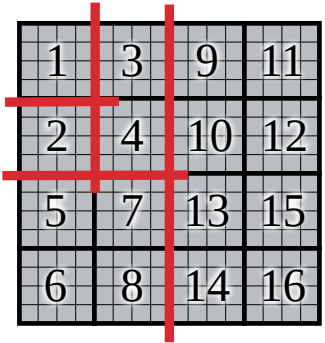
Caches facilitate data re-use

RAM is optimized for streaming

Tiling

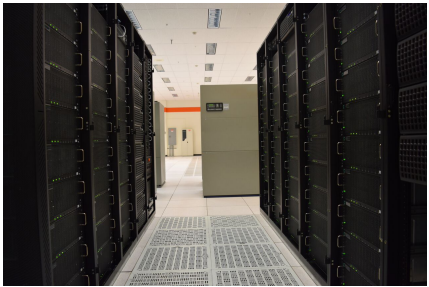


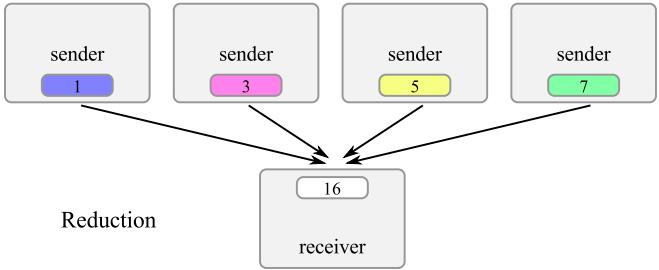
Cache-Oblivious Recursion



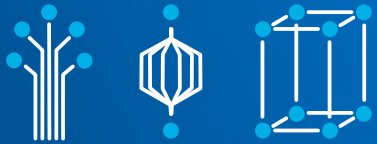
Lecture 5: Clusters and MPI

Clusters form distributed-memory systems with network interconnects





§6. Remote Access



Educational Computing Cluster