Thrill Tutorial: High-Performance Algorithmic Distributed Computing with C++

Timo Bingmann · 2019-08-29 @ gridka-school

Karlsruhe Institute of Technology

INSTITUTE OF THEORETICAL INFORMATICS – ALGORITHMICS

Abstract

In this tutorial we present our new distributed Big Data processing framework called Thrill. It is a C++ framework consisting of a set of basic scalable algorithmic primitives like mapping, reducing, sorting, merging, joining, and additional MPI-like collectives. This set of primitives can be combined into larger more complex algorithms, such as WordCount, PageRank, and suffix sorting. Such compounded algorithms can then be run on very large inputs using a distributed computing cluster with external memory.

After introducing the audience to Thrill we guide participants through the initial steps of downloading and compiling the software package. The tutorial then continues to give an overview of the challenges of programming real distributed machines and models and frameworks for achieving this goal. With these foundations, Thrill's DIA programming model is introduced with an extensive listing of DIA operations and how to actually use them. The participants are then given a set of small example tasks to gain hands-on experience with DIAs.

After the hands-on session, the tutorial continues with more details on how to run Thrill programs on clusters and how to generate execution profiles. Then, deeper details of Thrill's internal software layers are discussed to advance the participants' mental model of how Thrill executes DIA operations. The final hands-on tutorial is designed as a concerted group effort to implement K-means clustering for 2D points.

Table of Contents

1 Thrill Motivation Pitch

- Benchmarks and Introduction
- Tutorial: Clone, Compile, and Run Simple Example

Weak-Scaling Benchmarks

WordCountCC – $h \cdot 49\,\text{GiB}$ — 222 lines

- Reduce text files from CommonCrawl web corpus.

PageRank – $h \cdot 2.7\,\text{GiB}$, $|E| \approx h \cdot 158\,\text{M}$ — 410 lines

- Calculate PageRank using join of current ranks with outgoing links and reduce by contributions. 10 iterations.

TeraSort – $h \cdot 16\,\text{GiB}$ — 141 lines

- Distributed (external) sorting of 100 byte random records.

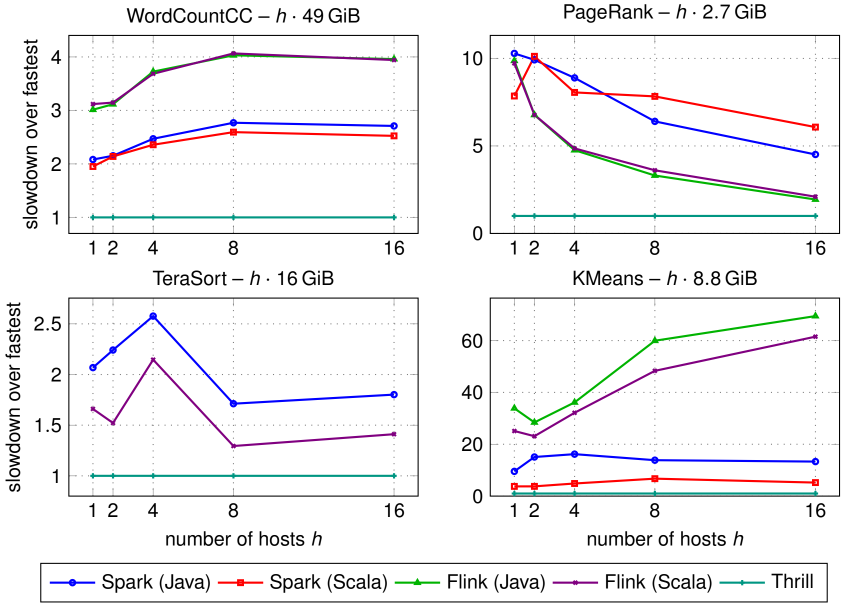
K-Means – $h \cdot 8.8\,\text{GiB}$ — 357 lines

- Calculate K-Means clustering with 10 iterations.

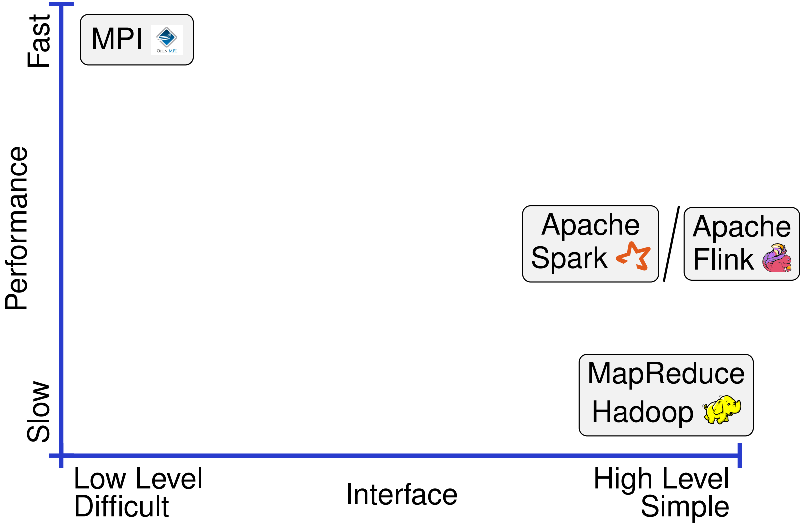
Platform: $h \times$ r3.8xlarge systems on Amazon EC2 Cloud

- 32 cores, Intel Xeon E5-2670v2, 2.5 GHz clock, 244 GiB RAM, 2 x 320 GB local SSD disk, $\approx$ 400 MiB/s read/write Ethernet network $\approx$ 1000 MiB/s throughput, Ubuntu 16.04.

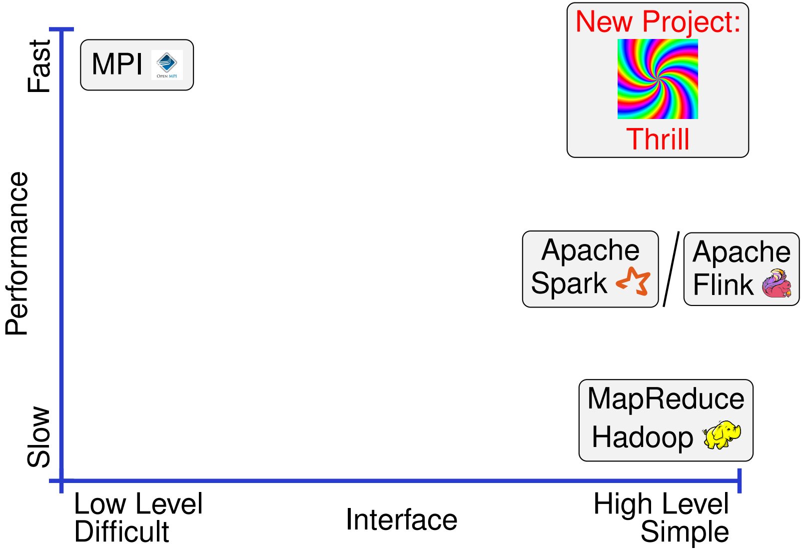
Experimental Results: Slowdowns



Big Data Batch Processing



Big Data Batch Processing



Big Data Batch Processing

Fast

Performance

Slow

MPI

New Project:

Thrill

Our Requirements:

- compound primitives into complex algorithms
- efficient simple data types,
- overlap computation and communication,
- automatic disk usage,
- C++, and in much more...

Apache Spark / Apache Flink

MapReduce Hadoop

Low Level
Difficult

Interface

High Level
Simple

Big Data Batch Processing

Fast

MPI

Lower Layers
of Thrill

New Project:
Thrill

Performance

Our Requirements:

- compound primitives into complex algorithms
- efficient simple data types,
- overlap computation and communication,
- automatic disk usage,
- C++, and in much more...

Apache Spark / Apache Flink

MapReduce
Hadoop

Slow

Low Level
Difficult

Interface

High Level
Simple

Thrill's Design Goals

- An easy way to program distributed algorithms in C++.
- Distributed arrays of small items (characters or integers).
- High-performance, parallelized C++ operations.
- Locality-aware, in-memory computation.
- Transparently use disk if needed
  $\Rightarrow$ external memory or cache-oblivious algorithms.
- Avoid all unnecessary round trips of data to memory (or disk).
- Optimize chaining of local operations.

Thrill is a moving target,
this tutorial is for the version in August 2019.

Thrill's Goal and Current Status

An easy way to program distributed external algorithms in C++.

Current Status:

- Open-source prototype at `http://github.com/thrill/thrill`.
- $\approx$ 60 K lines of C++14 code, written by $\geq$ 12 contributors.
- Published at IEEE Conference on Big Data [B, et al. '16]
- Faster than Apache Spark and Flink on five micro benchmarks: WordCount1000/CC, PageRank, TeraSort, and K-Means.

Case Studies:

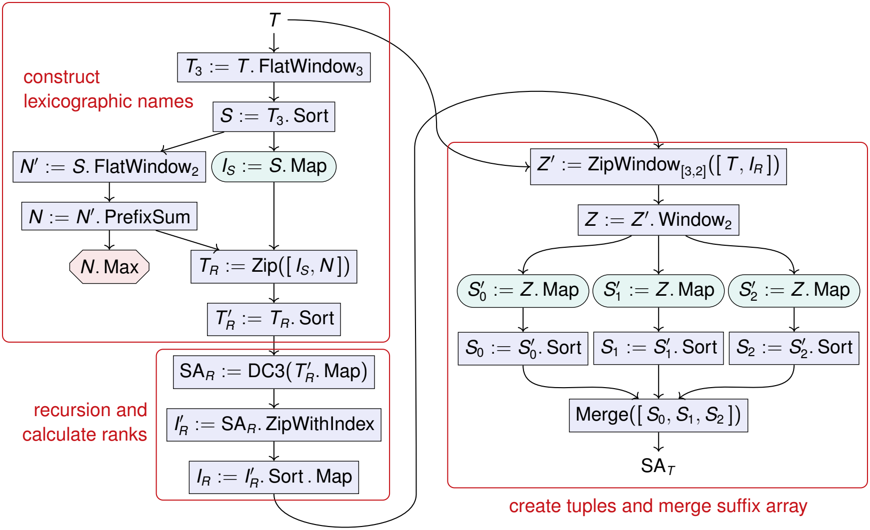
- Five suffix sorting algorithms [B, Gog, Kurpicz, BigData'18]
- Louvain graph clustering algorithm [Hamann et al. Euro-Par'18]
- Process scientific data on HPC (poster) [Karabin et al. SC'18]
- More: stochastic gradient descent, triangle counting, etc.
- Future: fault tolerance, scalability, predictability, and more.

Example: WordCount in Thrill

Karlsruhe Institute of Technology

```cpp
using Pair = std::pair<std::string, size_t>;
void WordCount(Context& ctx, std::string input, std::string output) {
    auto word_pairs = ReadLines(ctx, input)     // DIA<std::string>
    .FlatMap<Pair>(
        // flatmap lambda: split and emit each word
        [](const std::string& line, auto emit) {
            tlx::split_view(' ', line, [&](tlx::string_view sv) {
                emit(Pair(sv.to_string(), 1)); });
    });                                         // DIA<Pair>
    word_pairs.ReduceByKey(
        // key extractor: the word string
        [](const Pair& p) { return p.first; },
        // commutative reduction: add counters
        [](const Pair& a, const Pair& b) {
            return Pair(a.first, a.second + b.second);
    })                                          // DIA<Pair>
    .Map([](const Pair& p) {
        return p.first + ": " + std::to_string(p.second); })
    .WriteLines(output);                        // DIA<std::string>
}
```

Data-Flow Graph of DC3

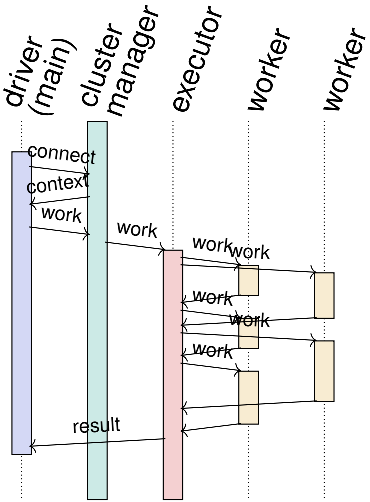


1 Thrill Motivation Pitch

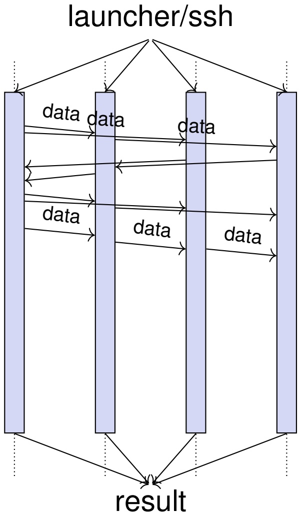
- Benchmarks and Introduction
- Tutorial: Clone, Compile, and Run Simple Example

Control Model: Spark vs. MPI/Thrill

Apache Spark



MPI and Thrill



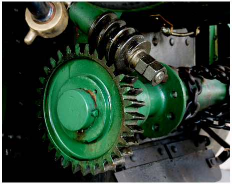
Tutorial: Clone, Compile, and Run

Karlsruhe Institute of Technology

This tutorial focuses on Linux and similar systems. Windows/Visual C++ is supported using CMake, but needs some extra steps.

- Clone the tutorial example repository:

```
git clone --recursive https://github.com/thrill/tutorial-project.git
```



- Compile with auto-detected C++14 GCC compiler:

```
$ cd tutorial-project
$ ./compile.sh
        -DTHRILL_BUILD_EXAMPLES=ON
```

- Run simple example:

```
$ cd build
$ ./simple
```

Tutorial: Run Simple Example

```cpp
#include <thrill/api/context.hpp>
#include <iostream>

void program(thrill::Context& ctx) {
    std::cout << "Hello World, I am "
              << ctx.my_rank() << std::endl;
}

int main(int argc, char* argv[]) {
    return thrill::Run(program);
}
```

Tutorial: Simple Example Output

```
Thrill: using 7.709 GiB RAM total, BlockPool=2.570 GiB,
    workers=657.877 MiB, floating=2.570 GiB.
Thrill: running locally with 2 test hosts and 4 workers per host
    in a local tcp network.
Thrill: using 7.709 GiB RAM total, BlockPool=2.570 GiB,
    workers=657.877 MiB, floating=2.570 GiB.
Thrill: no THRILL_LOG was found, so no json log is written.
[main 000000] FOXXLL v1.4.99 (prerelease/Release)
    (git a4a8aeee64743f845c5851e8b089965ea1c219d7)
[main 000001] foxxll: Using default disk configuration.
[main 000002] foxxll: Disk '/var/tmp/thrill.30713.tmp' is allocated,
    space: 1000 MiB, I/O implementation: syscall queue=0 devid=0 unlink_on_open
Hello World, I am 0
Hello World, I am 1
Hello World, I am 2
Hello World, I am 7
Hello World, I am 3
Hello World, I am 6
Hello World, I am 4
Hello World, I am 5
Thrill: ran 6.7e-05s with max 0.000 B in DIA Blocks, 0.000 B network traffic,
    0.000 B disk I/O, and 0.000 B max disk use.
malloc_tracker ### exiting, total: 1163264, peak: 1163264,
    current: 0 / 65536, allocs: 71, unfreed: 4
```

2 Introduction to Parallel Machines

- The Real Deal: Examples of Machines
- Networks: Types and Measurements
- Implementations and Frameworks

The Real Deal: HPC Supercomputers

Summit at Oak Ridge National Laboratory (ORNL)
#1 in TOP500 list since June 2018

4 356 nodes with two 22-core Power9 CPUs and six NVIDIA Tesla V100 GPUs each. The nodes are connected with a Mellanox dual-rail EDR InfiniBand network. In total 2 414 592 cores reaching 148.6 petaflops. 2×800 GB non-volatile RAM per node.

The Real Deal: HPC Supercomputers

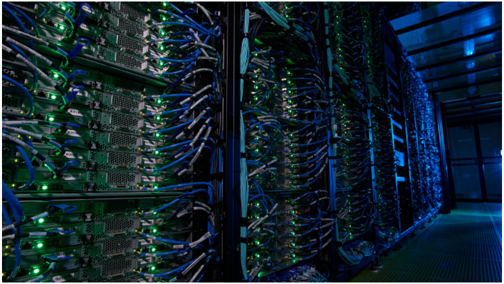
SuperMUC-NG at Leibniz Rechenzentrum (LRZ) in Munich
#9 in TOP500 list from June 2019

Picture: Veronika Hohenegger, LRZ

6 336 nodes with (24+24)-core Intel Xeon 8174 CPUs with 96 GiB RAM. The nodes are connected with an Intel Omni-Path 100 GB/s. In total 305 856 cores reaching 19.5 petaflops. No local disks.

The Real Deal: HPC Supercomputers

ForHLR II at Steinbuch Centre for Computing (SCC) at KIT



Close-up of ForHLR II, Andreas Drollinger, KIT (SCC)

1 152 nodes with two (10+10)-core Intel Xeon E5-2660 v3 with 64 GiB RAM. The nodes are connected with a Mellanox FDR adapter to an InfiniBand 4X EDR interconnect. In total 48 000 cores reaching about 1 petaflop. One 480 GB local SSD per node.

The Real Deal: Cloud Computing

Not much is public about their size, infrastructure, or even location.

Delivers virtualized computer, disk, and network resources.

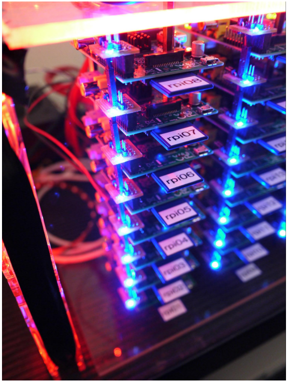
Probably built on commodity hardware, such as Intel processors, with some proprietary customizations and a virtualization stack.

Examples of AWS instances:

- m5.12xlarge has 48 vCPUs with 192 GB RAM and 10 Gb/s network, and costs $2.31 per hour
- i3.8xlarge has 32 vCPUs with 244 GB RAM, 10 Gb/s network, 4×1.9 TB NVMe SSDs, and costs $2.50 per hour

The Real Deal: Custom Local Clusters

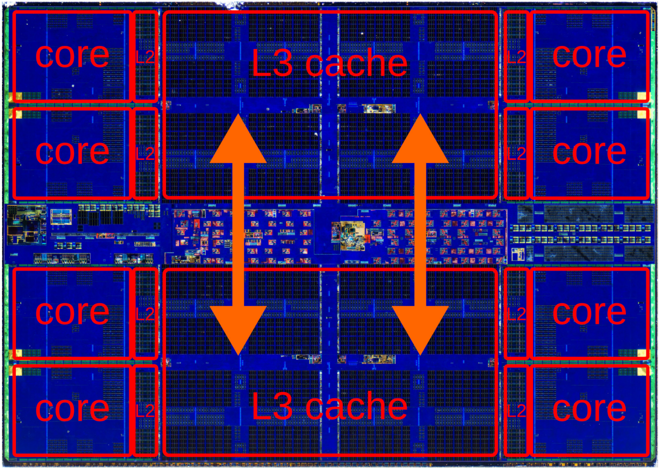
heterogeneous
server installations



Raspberry Pi clusters

photo and report by Joshua Kiepert, see
http://coen.boisestate.edu/ece/research-areas/raspberry-pi/

The Real Deal: Shared Memory



AMD Ryzen 5 3600, 6 cores, 3.60 GHz, 7 nm, 32 MiB L3 cache,
die photo from https://www.flickr.com/photos/130561288@N04/albums, modified

The Real Deal: GPUs

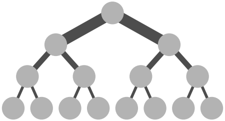
diagram from NVIDIA Tesla V100 GPU architecture whitepaper

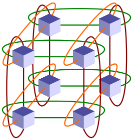
NVIDIA Tesla V100 with 80 streaming multiprocessors (SMs), each containing 64 CUDA cores, in total of 5 120 cores and up to 32 GB RAM.

2 Introduction to Parallel Machines

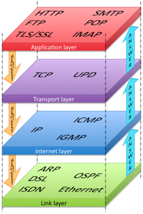
- The Real Deal: Examples of Machines
- Networks: Types and Measurements
- Implementations and Frameworks

Types of Networks



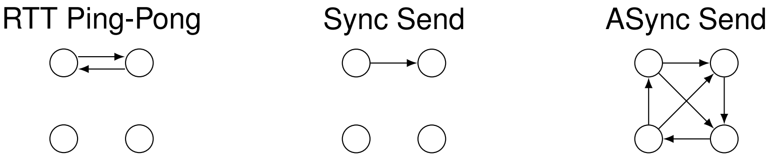


- HPC supercomputers:
  - remote direct memory access (RDMA)
  - different network topologies: fat trees, *k*D-torus, islands.
- cloud computing and local Ethernet clusters:
  - TCP/UDP/IP stack
  - switched 100 Mb/s, 1 Gb/s, 10 Gb/s, or more
- shared-memory many-core and GPU systems
  - implicit communication via cache coherence

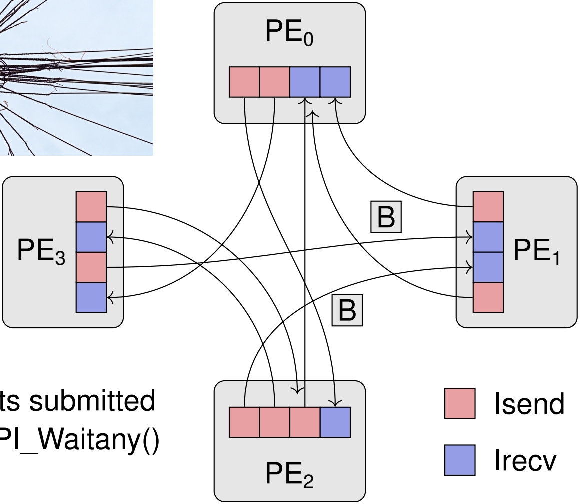


Round Trip Time (RTT) and Bandwidth

- 2 hosts in LAN at our institute at KIT — 2019-08-08
  RTT: 140 µs, bandwidth sync: 941 MiB/s
- 4 × r3.8xlarge AWS instances with 10 Gb/s net — 2016-07-14
  RTT: 100 µs, bandwidth sync: 389 MiB/s
- 4 × i3.4xlarge AWS instances with 10 Gb/s net — 2017-12-17
  RTT: 81 µs, bandwidth sync: 1 144 MiB/s, async: 4 278 MiB/s
- 4 × ForHLR II hosts with RDMA/4X EDR Infiniband — 2019-08-08
  RTT: 10.4 µs, bandwidth sync: 5 935 MiB/s, async: 5 554 MiB/s



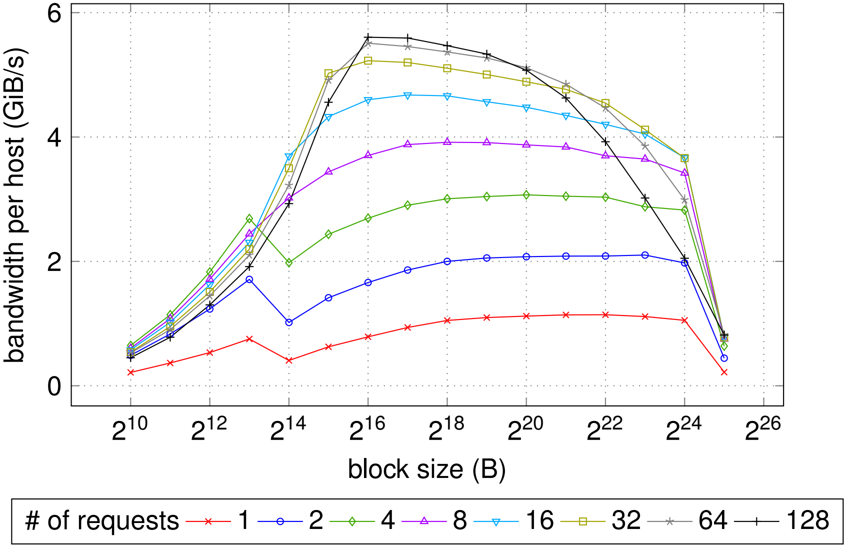
MPI Random Async Block Benchmark



requests submitted
with MPI_Waitany()

more: https://github.com/bingmann/mpi-random-block-test

Random Blocks on ForHLR II, 8 Hosts



Variety of Parallel Computing Hosts

- cluster types: homogeneous or heterogeneous
- host types: commodity hardware, virtual instances on cloud computing platforms, shared-memory many-core systems, GPUs, or HPC systems with RDMA.
- storage devices:
  - no local storage
  - local storage: rotational disks, SSD, or NVMe devices
  - transparent distributed storage
- network interconnect:
  - implicit communication protocols
  - explicit communication: Ethernet, virtual networking, RDMA/Infiniband, etc.

2 Introduction to Parallel Machines

- The Real Deal: Examples of Machines
- Networks: Types and Measurements
- Implementations and Frameworks

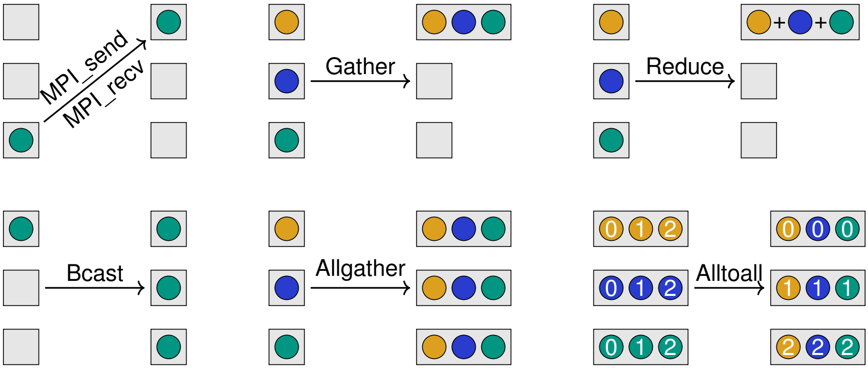
MPI (Message Passing Interface)

Karlsruhe Institute of Technology

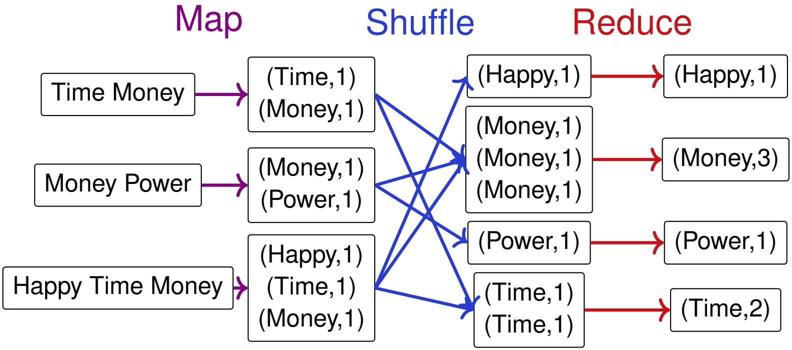
History:

- Version 1.0 from 1994 for C, C++, and Fortran.
- Still most used interface on supercomputers.

Collective Operations:



Map/Reduce Model



Computation model popularized in 2004
by Google with the name MapReduce.

Map/Reduce Framework

- Changes the perspective from the number of processors to how data is processed.
- A simple algorithmic and programming abstraction with
  - automatic parallelization of independent operations (map) and aggregation (reduce),
  - automatic distribution and balancing of data and work,
  - automatic fault tolerance versus hardware errors.

$\Rightarrow$ **all provided by MapReduce framework**

Apache Spark and Apache Flink

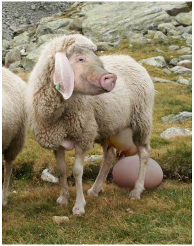
$$A \rightarrow \boxed{B := A.\mathrm{Map}()} \rightarrow \boxed{C := B.\mathrm{Sort}()} \rightarrow C$$

- New **post-**Map/Reduce frameworks use data-flow functional-style programming.
- Apache Spark started in 2009 in Berkeley.
  - central data structure: resilient distributed data sets (RDDs)
  - operations broken down into stages executed on cluster
  - driver initiates and controls execution of stages
- Apache Flink started as Stratosphere at TU Berlin.
  - first version (2010): “PACTs” and Nephele engine.
  - uses host language to construct data-flow graphs
  - optimizer and scheduler decide how to run them

Flavours of Big Data Frameworks

Karlsruhe Institute of Technology

- Batch Processing
  Google's MapReduce, Hadoop MapReduce, Apache Spark, Apache Flink (Stratosphere).
- High Performance Computing (Supercomputers)
  MPI
- Real-time Stream Processing
  Apache Storm, Apache Spark Streaming.
- Interactive Cached Queries
  Google's Dremel, Powerdrill and BigQuery, Apache Drill.
- Sharded (NoSQL) Databases and Data Warehouses
  MongoDB, Apache Cassandra, Google BigTable, Amazon RedShift.
- Graph Processing
  Google's Pregel, GraphLab, Giraph, GraphChi.
- Machine Learning Frameworks and Libraries
  Tensorflow, Keras, scikit-learn, Microsoft Cognitive Toolkit.



eierlegende Wollmilchsau
CC BY-SA Georg Mittenecker

3 The Thrill Framework

Big Data Batch Processing

Fast

Performance

Slow

MPI

New Project:

Thrill

Our Requirements:

- compound primitives into complex algorithms
- efficient simple data types,
- overlap computation and communication,
- automatic disk usage,
- C++, and in much more...

Apache Spark / Apache Flink

MapReduce Hadoop

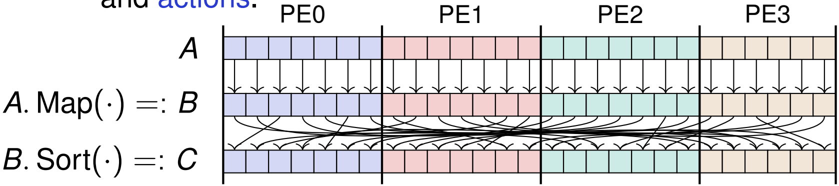
Low Level
Difficult

Interface

High Level
Simple

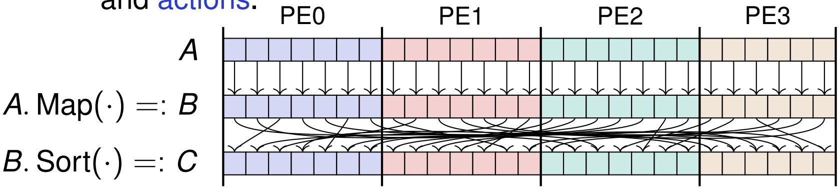
Distributed Immutable Array (DIA)

- User Programmer's View:
  - `DIA<T>` = result of an operation (local or distributed).
  - Model: distributed array of items T on the cluster
  - Cannot access items directly, instead use transformations and actions.



Distributed Immutable Array (DIA)

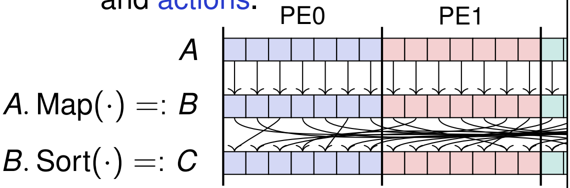
- User Programmer's View:
  - `DIA<T>` = result of an operation (local or distributed).
  - Model: distributed array of items T on the cluster
  - Cannot access items directly, instead use transformations and actions.

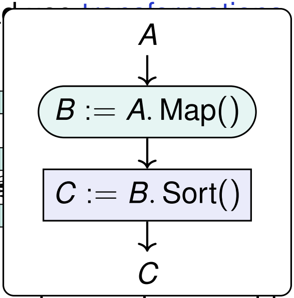


- Framework Designer's View:
  - Goals: distribute work, optimize execution on cluster, add redundancy where applicable. $\Longrightarrow$ build data-flow graph.
  - `DIA<T>` = chain of computation items
  - Let distributed operations choose "materialization".

Distributed Immutable Array (DIA)

- User Programmer's View:
  - `DIA<T>` = result of an operation (local or distributed).
  - Model: distributed array of items T on the cluster
  - Cannot access items directly, instead
    and actions.





- Framework Designer's View:
  - Goals: distribute work, optimize execution on cluster, add redundancy where applicable. $\Longrightarrow$ build data-flow graph.
  - `DIA<T>` = chain of computation items
  - Let distributed operations choose “materialization”.

List of Primitives (Excerpt)

- Local Operations (LOp): input is one item, output ≥ 0 items.
  Map(), Filter(), FlatMap().
- Distributed Operations (DOp): input is a DIA, output is a DIA.

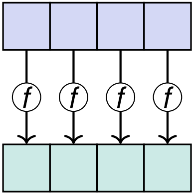
Sort()	Sort a DIA using comparisons.
ReduceBy()	Shuffle with Key Extractor, Hasher, and associative Reducer.
GroupBy()	Like ReduceBy, but with a general Reducer.
PrefixSum()	Compute (generalized) prefix sum on DIA.
Window_k()	Scan all k consecutive DIA items.
Zip()	Combine equal sized DIAs item-wise.
Union()	Combine equal typed DIAs in arbitrary order.
InnerJoin()	Join items from two DIAs by key.

- Actions: input is a DIA, output: ≥ 0 items on every worker.
  Sum(), Min(), ReadLines(), WriteLines(), pretty much open.

Local Operations (LOps)

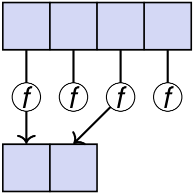
Map$(f) : \langle A \rangle \to \langle B \rangle$

$f : A \to B$



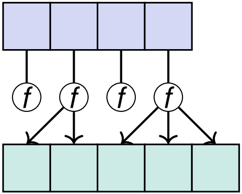
Filter$(f) : \langle A \rangle \to \langle A \rangle$

$f : A \to \{\mathit{false}, \mathit{true}\}$



FlatMap$\langle B \rangle (f) : \langle A \rangle \to \langle B \rangle$

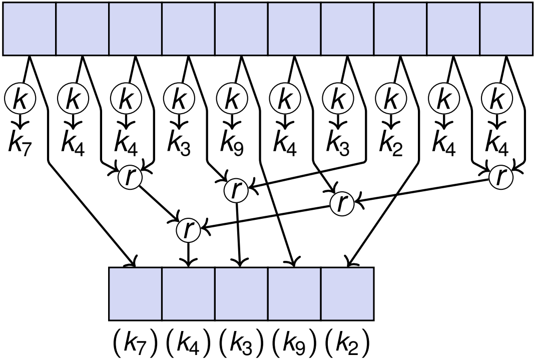
$f : A \to \text{array}(B)$



DOps: ReduceByKey

$$\textbf{ReduceByKey}(k, r) : \langle A \rangle \to \langle A \rangle$$

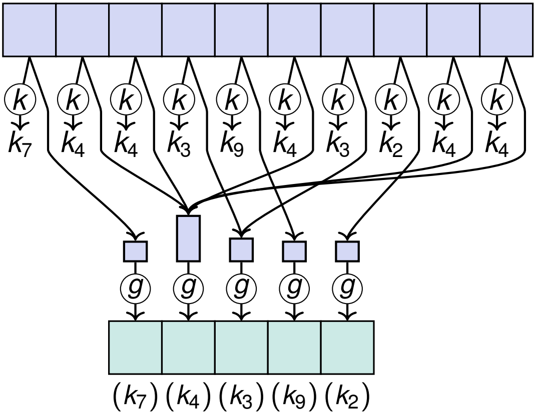
$k : A \to K$	key extractor
$r : A \times A \to A$	reduction



DOps: GroupByKey

$$\textbf{GroupByKey}(k, g) : \langle A \rangle \to \langle B \rangle$$

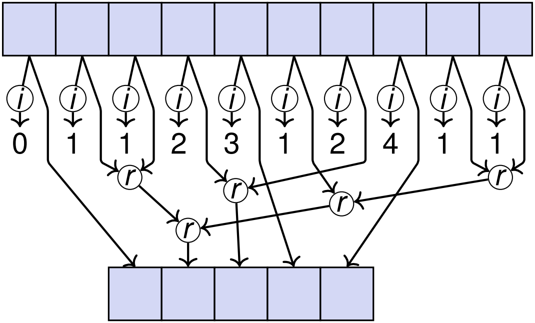
$k : A \to K$	key extractor
$g : iterable(A) \to B$	group function



DOps: ReduceToIndex

$$\textbf{ReduceToIndex}(i, n, r) : \langle A \rangle \to \langle A \rangle$$

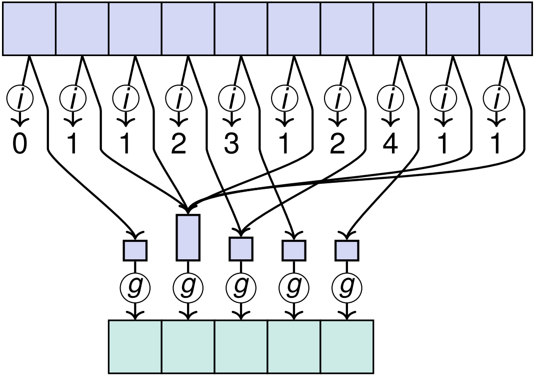
$i : A \to \{0..n-1\}$	index extractor
$n \in \mathbb{N}_0$	result size
$r : A \times A \to A$	reduction



DOps: GroupToIndex

$$\textbf{GroupToIndex}(i, n, g) : \langle A \rangle \to \langle B \rangle$$

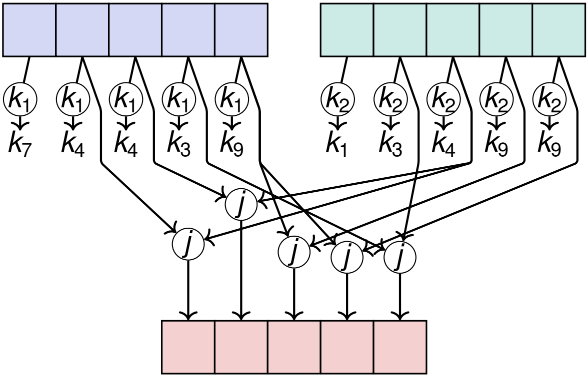
$i : A \to \{0..n-1\}$	index extractor
$n \in \mathbb{N}_0$	result size
$g : \mathit{iterable}(A) \to B$	group function



DOps: InnerJoin

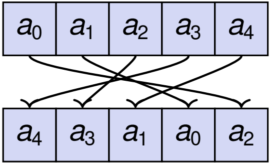
$$\textbf{InnerJoin}(k_1, k_2, j) : \langle A \rangle \times \langle B \rangle \to \langle C \rangle$$

$k_1 : A \to K$	key extractor for A
$k_2 : B \to K$	key extractor for B
$j : A \times B \to C$	join function

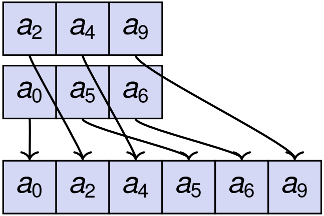


DOps: Sort and Merge

Sort$(o) : \langle A \rangle \to \langle A \rangle$
$o : A \times A \to \{\mathit{false}, \mathit{true}\}$
(less) order relation



Merge$(o) : \langle A \rangle \times \langle A \rangle \cdots \to \langle A \rangle$
$o : A \times A \to \{\mathit{false}, \mathit{true}\}$
(less) order relation

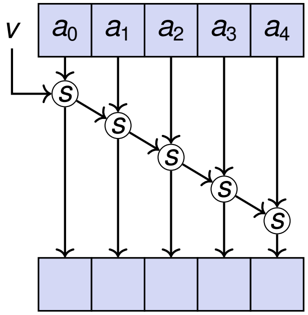


DOps: PrefixSum and ExPrefixSum

PrefixSum$(s, v) : \langle A \rangle \to \langle A \rangle$

$s : A \times A \to A$ sum function

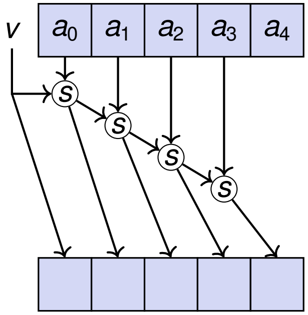
$v : A$ initial value



ExPrefixSum$(s, v) : \langle A \rangle \to \langle A \rangle$

$s : A \times A \to A$ sum function

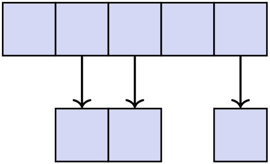
$v : A$ initial value

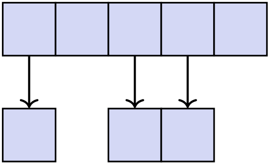


Sample (LOp), BernoulliSample (DOp)

Sample(k) : $\langle A \rangle \to \langle A \rangle$
$k \in \mathbb{N}_0$ result size

BernoulliSample(p) : $\langle A \rangle \to \langle A \rangle$
$p \in [0, 1]$ probability



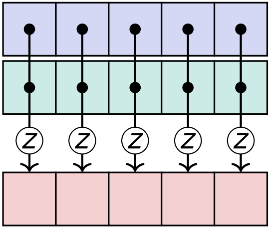


DOps: Zip and Window

Zip$(z) : \langle A \rangle \times \langle B \rangle \cdots \to \langle C \rangle$

$z : A \times B \to C$

zip function



ZipWithIndex$(z) : \langle A \rangle \to \langle C \rangle$

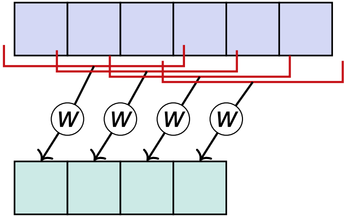
$z : A \times \mathbb{N}_0 \to C$

zip function

Window$(k, w) : \langle A \rangle \to \langle B \rangle$

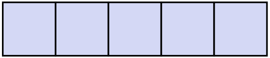
$k \in \mathbb{N}$ window size

$w : A^k \to B$ window function



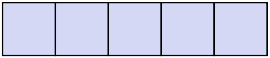
Special Ops: Cache and Collapse

Cache() : $\langle A \rangle \to \langle A \rangle$



Materializes a DIA,
needed e.g. for caching or
random data generation.

Collapse() : $\langle A, f_1, f_2 \rangle \to \langle A \rangle$



Folds local operation
lambdas f_1, f_2 into a DIA,
needed for iterations.

Source DOps: Generate, -ToDIA

Generate$(n, g) : \langle A \rangle$

$n \in \mathbb{N}_0$ result size

$g : \{0..n-1\} \to A$ generator

$g(0)$	$g(1)$	$g(2)$	$g(3)$	$g(4)$

Generate$(n) : \langle \mathbb{N}_0 \rangle$

$n \in \mathbb{N}_0$ result size

0	1	2	3	4

ConcatToDIA$(v) : \langle A \rangle$

$v : vector(A)$ input data

$v[0]$	$v[1]$	$v[0]$	$v[1]$	$v[2]$	$v[0]$

PE0 PE1 PE2

EqualToDIA$(v) : \langle A \rangle$

$v : vector(A)$ input data

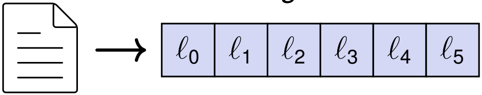
$v[0]$	$v[1]$	$v[2]$	$v[3]$	$v[4]$	$v[5]$

PE0 PE1 PE2

Source DOps: ReadLines, ReadBinary

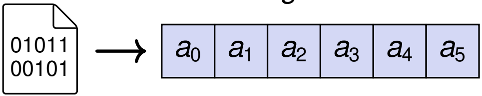
ReadLines(f) : $\langle$`std::string`$\rangle$

f : *string* list of files



ReadBinary$\langle A\rangle(f)$: $\langle A\rangle$

f : *string* list of files



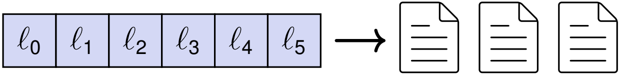
Items A are serialized in Thrill's binary representation.

Both either read from a common distributed file system (DFS), or concatenate from all PEs with the "local-storage" flag.

Actions: WriteLines, WriteBinary

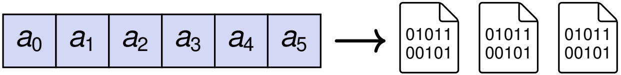
WriteLines(f) : $\langle$`std::string`$\rangle \rightarrow$ *void*

f : *string* path/file pattern



WriteBinary(f) : $\langle A\rangle \rightarrow$ *void*

f : *string* path/file pattern



Items A are serialized with Thrill's binary representation.

Each PE writes one or more files to the DFS or local disk.

Actions: Size, Print, and more

$\textbf{Size}() : \langle A \rangle \to \mathbb{N}_0$

$[\ \ |\ \ |\ \ |\ \ ] \longrightarrow 4$

$\textbf{AllGather}() : \langle A \rangle \to \mathit{vector}(A)$

$[\, a_0 \mid a_1 \mid a_2 \mid a_3 \,]$

$\longrightarrow [\, a_0, a_1, a_2, a_3 \,]$

$\textbf{Print}(t) : \langle A \rangle \to \mathit{void}$

$t : \mathit{string}$ variable name

$\textbf{Gather}(t) : \langle A \rangle \to \mathit{vector}(A)$

$t \in \mathbb{N}_0$ target worker

$\textbf{Execute}() : \langle A \rangle \to \mathit{void}$

$\textbf{Sum}(s) : \langle A \rangle \to A$

$s : A \times A \to A$ sum function

$[\, a_0 \mid a_1 \mid a_2 \mid a_3 \,] \longrightarrow \sum_{i=0}^{3} a_i$

also: **Min**() and **Max**().

$\textbf{AllReduce}(s) : \langle A \rangle \to \mathit{vector}(A)$

$s : A \times A \to A$ sum function

$[\, a_0 \mid a_1 \mid a_2 \mid a_3 \,]$

$\longrightarrow s(s(s(a_0, a_1), a_2), a_3)$

3 The Thrill Framework

Playing with DIA Operations

How to get from the illustrated DIA operation to C++ code:

- Many operations have multiple variants and more parameters.
- The Doxygen documentation contains a very technical but complete list of DIA operations:

  https://project-thrill.org/docs/master/group___dia___api.html

Distributed Operations (DOps)

DIA API Operations

Modules

Modules

Free Operation Functions

Distributed Operations (DOps)

This list of DOps are methods of the **main DIA class** and called as `A.Method(params)`. Methods combining two or more DIAs are available as **free functions**.

```
template<typename KeyExtractor , typename ReduceFunction , typename ReduceConfig = class DefaultReduceConfig>
            auto  ReduceByKey (const KeyExtractor &key_extractor, const ReduceFunction &reduce_function, const ReduceConfig &reduce_config=Re
```

ReduceByKey is a DOp, which groups elements of the DIA with the key_extractor and reduces each key-bucket to a single element using the associative reduce_func

```
template<typename KeyExtractor , typename ReduceFunction , typename ReduceConfig , typename KeyHashFunction >
            auto  ReduceByKey (const KeyExtractor &key_extractor, const ReduceFunction &reduce_function, const ReduceConfig &reduce_config, c
```

ReduceByKey is a DOp, which groups elements of the DIA with the key_extractor and reduces each key-bucket to a single element using the associative reduce_func

```
template<typename KeyExtractor , typename ReduceFunction , typename ReduceConfig , typename KeyHashFunction , typename KeyEqualFunction >
            auto  ReduceByKey (const KeyExtractor &key_extractor, const ReduceFunction &reduce_function, const ReduceConfig &reduce_config, c
```

ReduceByKey is a DOp, which groups elements of the DIA with the key_extractor and reduces each key-bucket to a single element using the associative reduce_func

```
template<bool VolatileKeyValue, typename KeyExtractor , typename ReduceFunction , typename ReduceConfig = class DefaultReduceConfig, typename KeyHashFunction = std::hash<ty
```

Playing with DIA Operations

How to get your first DIA object:

- To use a DIA operation (CamelCase), include the corresponding header (snake_case): ReduceByKey $\rightarrow$ `reduce_by_key.hpp`.
- Initial DIAs are created from Source operations. These require the `thrill::Context` as first parameter.

```
#include <thrill/api/context.hpp>
#include <thrill/api/read_lines.hpp>

void program(thrill::Context& ctx) {
    auto lines = ReadLines(ctx, "/etc/hosts");
    lines.Print("lines");
}
int main(int argc, char* argv[]) {
    return thrill::Run(
        [&](thrill::Context& ctx) { return program(ctx); });
}
```

Playing with DIA Operations

Applying operations to DIA objects:

- DIA objects have many operations like `.Sum()` as methods, but there are also free functions like `Zip()` and `ReadLines()`.
- Generally use auto instead of `DIA<T>`:

```
#include <thrill/api/read_lines.hpp>
#include <thrill/api/size.hpp>

void program(thrill::Context& ctx) {
    auto lines = ReadLines(ctx, "/etc/hosts");
    std::cout << "lines: " << lines.Size() << std::endl;
}
```

- Or use chaining of operations:

```
void program(thrill::Context& ctx) {
    size_t num_lines = ReadLines(ctx, "/etc/hosts").Size();
    std::cout << "lines: " << num_lines << std::endl;
}
```

Playing with DIA Operations

More advanced uses of DIA objects:

- DIA objects are only handles to actual graphs nodes in the DIA data-flow. This means they are copied as references.
- It is straight-forward to have functions with DIAs as parameters and return type. Again, prefer templates and the auto keyword.

```
template <typename InputDIA>
auto LinesToLower(const InputDIA& input_dia) {
    return input_dia.Map(
        [](const std::string& line) {
            return tlx::to_lower(line);
        });
}
void program(thrill::Context& ctx) {
    auto lines = ReadLines(ctx, "/etc/hosts");
    std::cout << "lines: " << LinesToLower(lines).Size() << "\n";
}
```

Playing with DIA Operations

- Use C++11 lambdas for functor parameters.

```
using Pair = std::pair<std::string, size_t>;
void program(thrill::Context& ctx) {
    ReadLines(ctx, "/etc/hosts")
    .FlatMap<Pair>(
        // flatmap lambda: split and emit each word
        [](const std::string& line, auto emit) {
            tlx::split_view(' ', line, [&](tlx::string_view sv) {
                emit(Pair(sv.to_string(), 1)); });
    })
    .ReduceByKey(
        // key extractor: the word string
        [](const Pair& p) { return p.first; },
        // commutative reduction: add counters
        [](const Pair& a, const Pair& b) {
            return Pair(a.first, a.second + b.second);
    })
    .Execute();
}
```

Context Methods for Synchronization

The `Context` object also has many useful methods:

- `ctx.my_rank()` – rank of current worker thread.

```
      if (ctx.my_rank() == 0)
          std::cout << "lines: " << num_lines << '\n';
```

  also: `host_rank()`, `num_hosts()`, `num_workers()`.
- `y = ctx.net.Broadcast(x, 0);`
  MPI-style broadcast of x from worker 0 as y on all.
- `y = ctx.net.PrefixSum(x);`
  MPI-style prefix-sum of x with result y. also: `ExPrefixSum`.
- `y = ctx.net.AllReduce(x);`
  MPI-style all-reduce of x with result y. also: `Reduce`.
- `ctx.net.Barrier();`
  MPI-style synchronization barrier

Serializing Objects in DIAs

Thrill needs serialization methods for objects in DIAs.

- automatically supported are:
  - All plain old data types (PODs) (except pointers), which are plain integers, characters, doubles, and fixed-length structs containing such.
  - `std::string`, `std::pair`, `std::tuple`, `std::vector`, and `std::array`, if the contained type is serializable.
- otherwise, add a `serialize()` method:

```
#include <thrill/data/serialization_cereal.hpp>
struct Item {
    std::string string;
    size_t      value;
    template <typename Archive>
    void serialize(Archive& ar) {
        ar(string, value);
    }
};
```

Tutorial: Playing with DIAs

Hands-on Tutorial Part

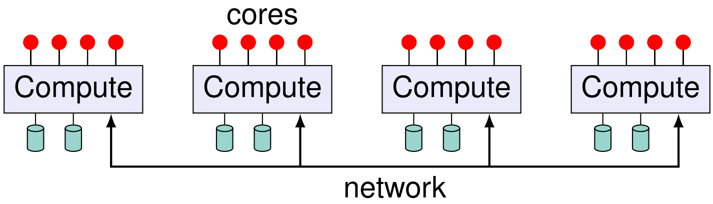
Objective:
Write and run some simple programs using DIA operations.

Some Ideas/Tasks:

- Read a text file, sort the lines, and write the result.
- Read a text file, transform all lines to lower case, and write them.
- Read a text file and calculate the average line length.
- Read a binary file as characters and count how many of each character occurs. Tip: use ReduceToIndex.
- Calculate the top 100 words in a text file and output all lines in which they occur.

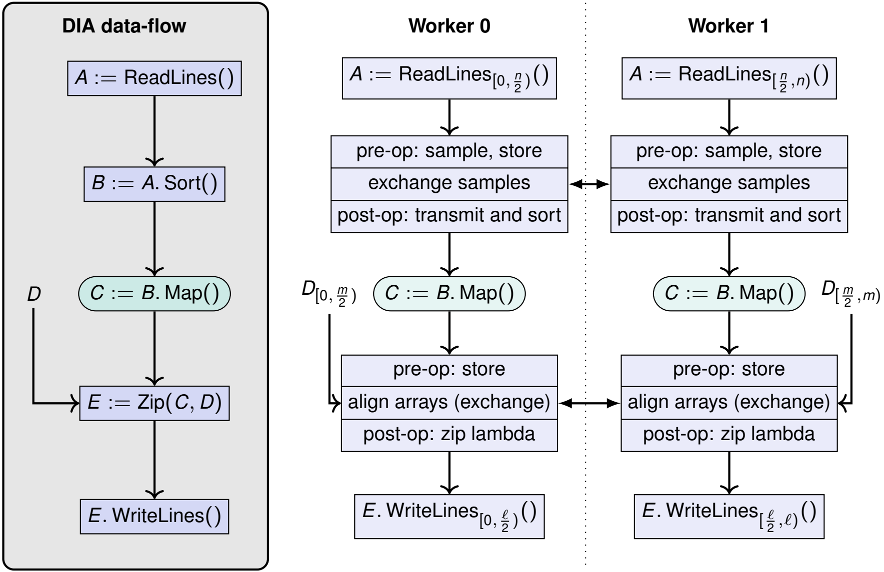
3 The Thrill Framework

Execution on Cluster

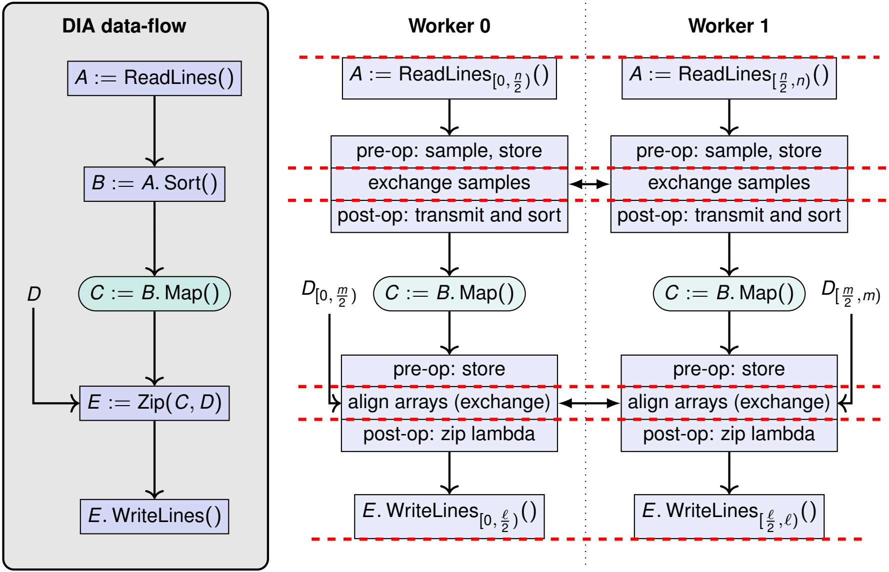


- Compile program into one binary, running on all hosts.
- Collective coordination of work on compute hosts, like MPI.
- Control flow is decided on by using C++ statements.
- Runs on MPI HPC clusters and on Amazon's EC2 cloud.

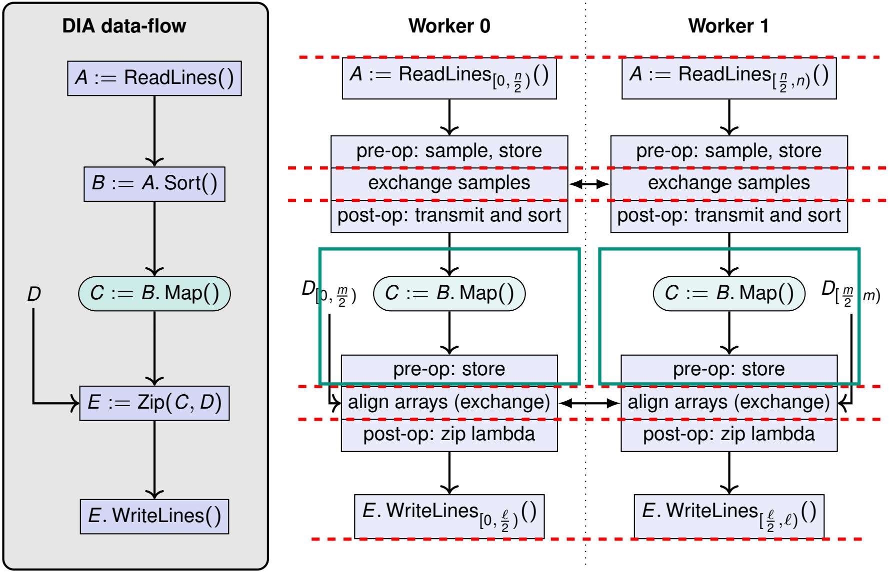
Mapping Data-Flow Nodes to Cluster



Mapping Data-Flow Nodes to Cluster



Mapping Data-Flow Nodes to Cluster



3 The Thrill Framework

- Thrill's DIA Abstraction and List of Operations
- Tutorial: Playing with DIA Operations
- Execution of Collective Operations in Thrill
- Tutorial: Running Thrill on a Cluster
- Tutorial: Logging and Profiling
- Going Deeper into Thrill
- Tutorial: First Steps towards K-Means
- Conclusion

Tutorial: Running Thrill on a Cluster

Supported Network Systems and Launchers:

- single multi-core machine
- cluster with ssh access and TCP/IP network
- MPI as startup system and transport network

Goal is to launch a Thrill binary on all hosts and pass information on how to contact the others.

Thrill reads environment variables for configuration.
(Configuration files would have to be copied to all hosts.)

Tutorial: On One Multi-Core Machine

- This is the default startup mode for easy development.
  You have already used it:

```
Thrill: running locally with 2 test hosts and 4 workers per host
    in a local tcp network.
```

- Default local settings are to split the cores on the machine into two virtual hosts, which communicate via local TCP sockets.
- Options to change the default settings:
  - `THRILL_LOCAL`: number of virtual hosts
  - `THRILL_WORKERS_PER_HOST`: workers per host

Tutorial: Running via ssh

- Mode for plain Linux machines connected via TCP/IP.
- a) Install ssh keys on all machines for password-less login.
- b) use `thrill/run/ssh/invoke.sh` script with
  - `-h "host1 host2 host3"` (host list)
  - `-u u1234` (remote user name)
  - `thrill-binary` (binary and arguments)
- two setups:
  - with a common file system (NFS, ceph, Lustre, etc)
    $\Rightarrow$ simply call the binary
  - without common file system (stand-alone machines).
    $\Rightarrow$ add `-c` to copy the binary to all hosts.

Tutorial: Running via MPI

- For running on HPC clusters, Thrill can use MPI directly. MPI is auto-detected, no configuration is needed.

- Check that `cmake` finds the MPI libraries when compiling:

```
-- Found MPI_C: /usr/lib64/libmpi.so (found version "3.1")
-- Found MPI_CXX: /usr/lib64/libmpi_cxx.so (found version "3.1")
-- Found MPI: TRUE (found version "3.1")
```

- Run with `mpirun`:

```
mpirun -H "host1,host2" thrill-binary
```

- On HPC clusters: use SLURM to launch with MPI, use only one task per host.

Tutorial: Environment Variables

- `THRILL_RAM` e.g. `=16GiB`
  override the maximum amount of RAM used by Thrill
- `THRILL_WORKERS_PER_HOST`
  override the number of workers per host
- `THRILL_LOG` e.g. `=out`
  write log and profile to JSON file, e.g. "`out-host-123.json`".

Environment variables can be set

- directly: `THRILL_RAM=16GiB program`
- with invoke.sh: `THRILL_RAM=16GiB invoke.sh program`
- or by mpirun: `mpirun -x THRILL_RAM=16GiB program`

3 The Thrill Framework

- Thrill's DIA Abstraction and List of Operations
- Tutorial: Playing with DIA Operations
- Execution of Collective Operations in Thrill
- Tutorial: Running Thrill on a Cluster
- Tutorial: Logging and Profiling
- Going Deeper into Thrill
- Tutorial: First Steps towards K-Means
- Conclusion

Tutorial: Logging and Profiling

- Thrill contains a built-in logging and profiling mechanism.
- To activate: set the environment variable `THRILL_LOG=abc`.
- Thrill writes logs to `abc-host0.json` in a JSON format.
- Use the included tool `json2profile` to generate HTML graphs.

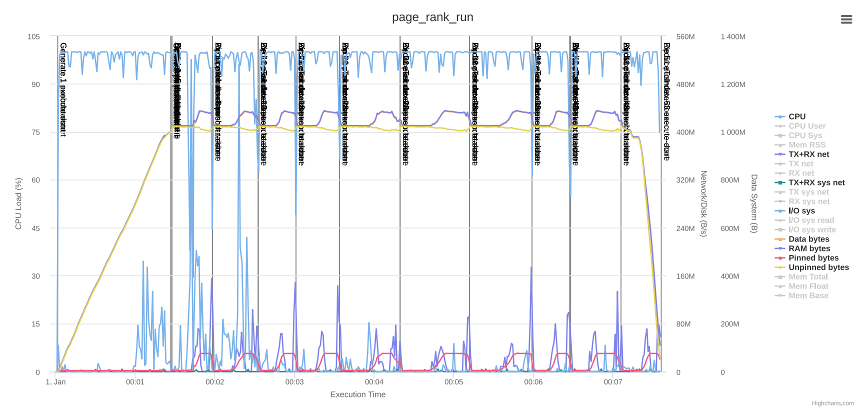
For example[1]:

```
$ cd ~/thrill/build/examples/page_rank/
$ THRILL_LOG=ourlog ./page_rank_run --generate 100000
$ ls -la ourlog*
(this should show ourlog-host0.json and ourlog-host1.json)
$ ~/thrill/build/misc/json2profile ourlog*.json > profile.html
```

And then visit `profile.html` with a browser.

[1] (adapt paths if in tutorial-project)

Tutorial: Example Profile



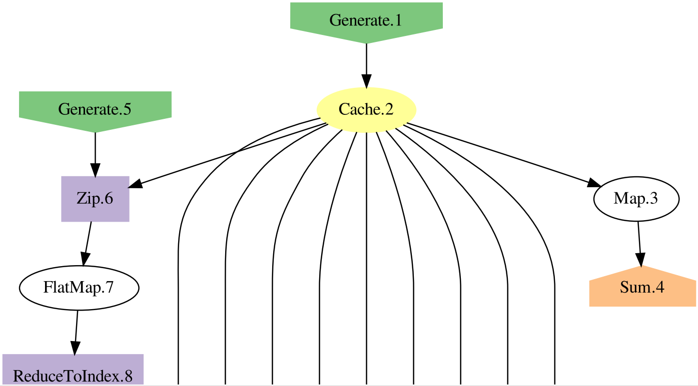
Summary

Running time	456.467 s	
CPU user+sys average	97.3114 %	
CPU user average	92.1755 %	
TX+RX net total	376.272 MiB	394549930 B
TX net total	7.062 MiB	7405436 B
RX net total	369.210 MiB	387144494 B

Tutorial: Output DIA Data-Flow Graph

- The DIA data-flow graph can also be extracted and automatically drawn with `dot` from the JSON log file:

```
$ ~/thrill/misc/json2graphviz.py ourlog-host-0.json > page_rank.dot
$ dot -Tps -o page_rank.ps page_rank.dot
or
$ dot -Tsvg -o page_rank.svg page_rank.dot
```



3 The Thrill Framework

Layers of Thrill

Karlsruhe Institute of Technology

api: High-level User Interface
DIA<T>, Map, FlatMap, Filter, Reduce, Sort, Merge, ...

core: Internal Algorithms
reducing hash tables (bucket and linear probing), multiway merge, stage executor

vfs: Data FS
local, S3, HDFS

data: Data Layer
Block, File, BlockQueue, Reader, Writer, Multiplexer, Streams, BlockPool (paging)

io: Async File I/O
borrowed from STXXL

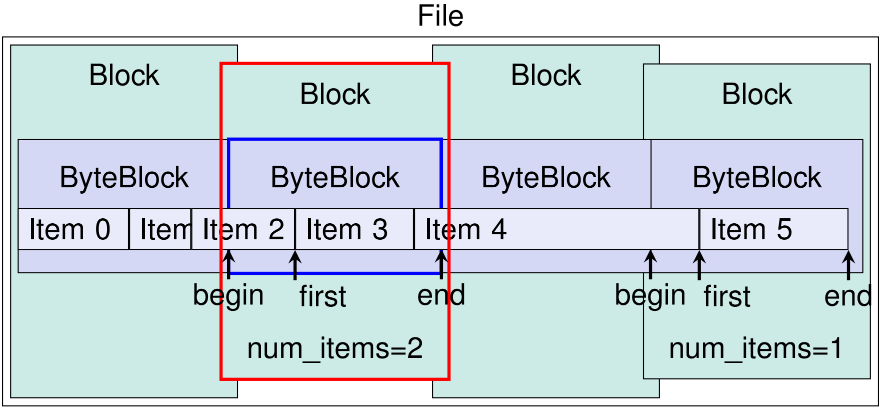
net: Network Layer
(Binomial Tree) Broadcast, Reduce, AllReduce, Async-Send/Recv, Dispatcher
Backends:

mock	tcp	mpi

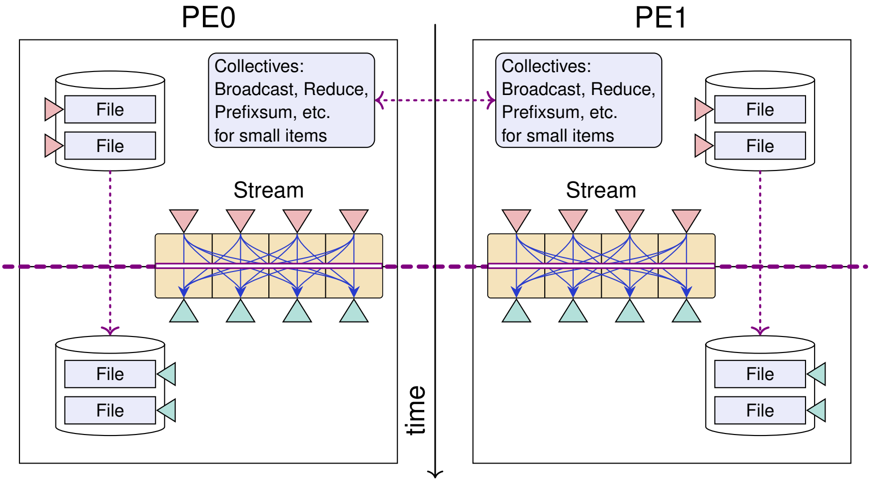
common: Common Tools
Logger, Delegates, Math, ...

mem: Memory Limitation
Allocators, Counting

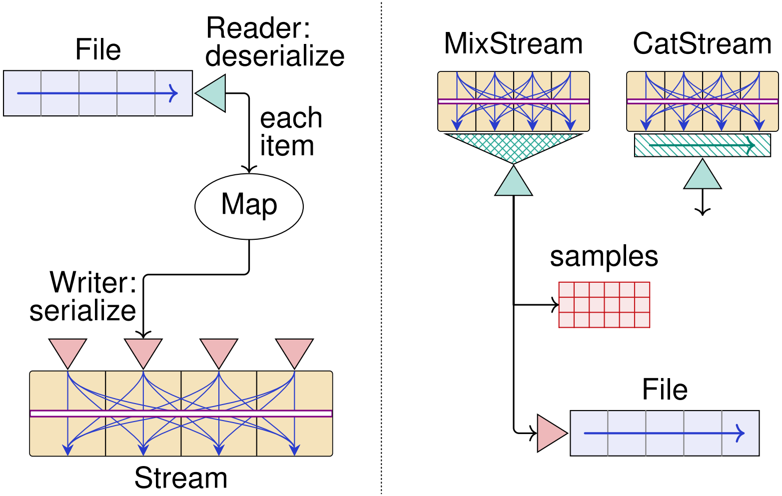
File – Variable-Length C++ Item Store



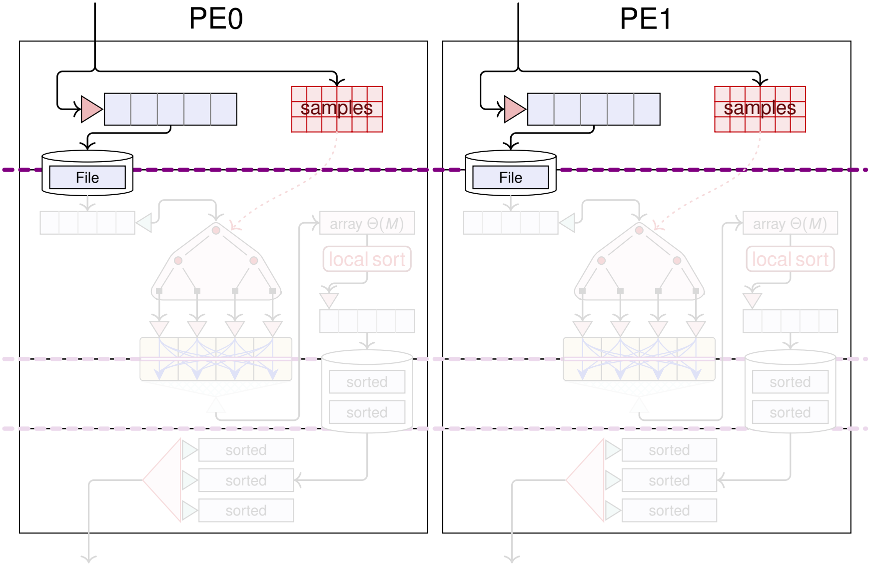
Thrill's Communication Abstraction



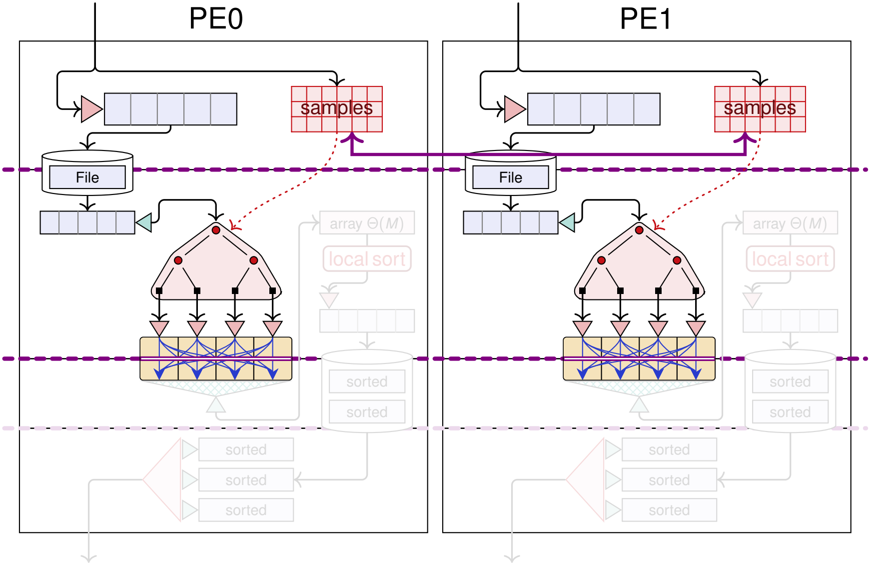
Thrill's Data Processing Pipelines



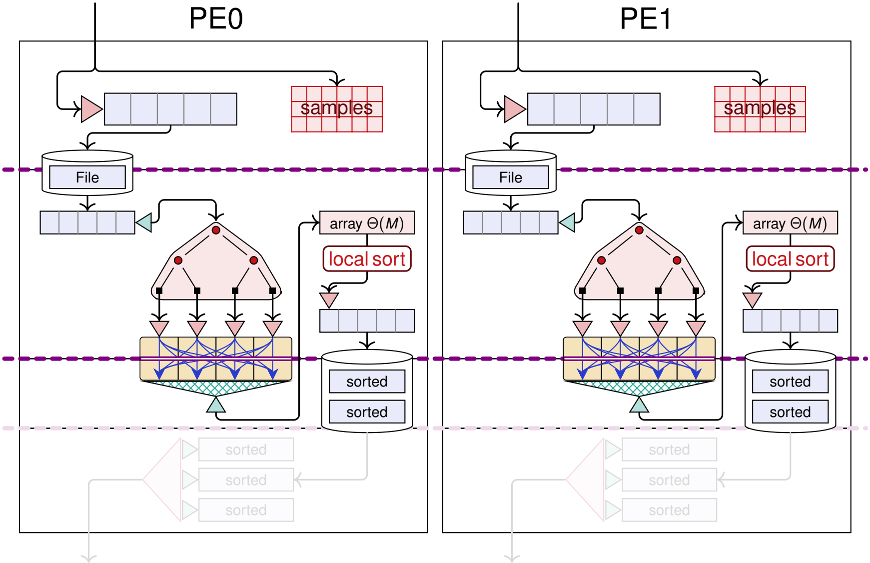
Thrill's Current Sample Sort



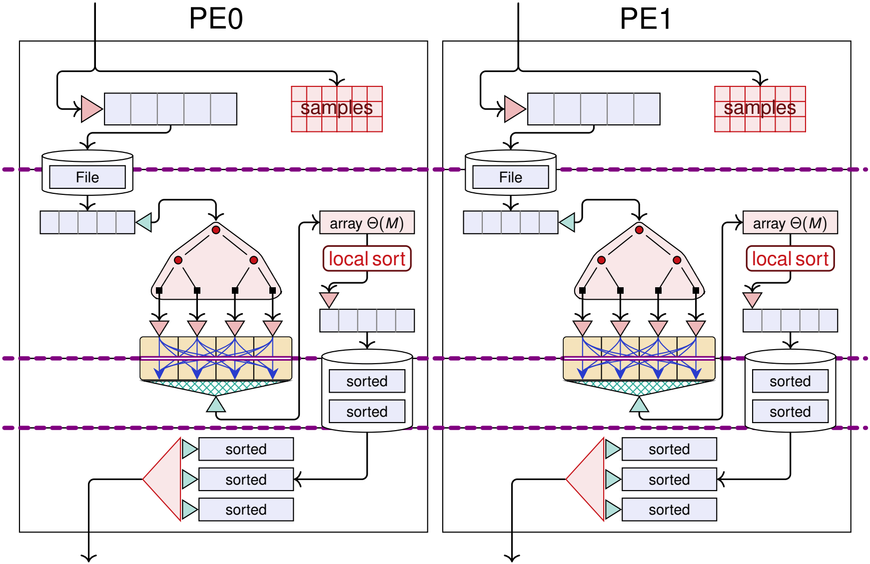
Thrill's Current Sample Sort



Thrill's Current Sample Sort



Thrill's Current Sample Sort

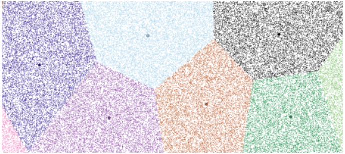


3 The Thrill Framework

- Thrill's DIA Abstraction and List of Operations
- Tutorial: Playing with DIA Operations
- Execution of Collective Operations in Thrill
- Tutorial: Running Thrill on a Cluster
- Tutorial: Logging and Profiling
- Going Deeper into Thrill
- Tutorial: First Steps towards K-Means
- Conclusion

Tutorial: First Steps towards *k*-Means

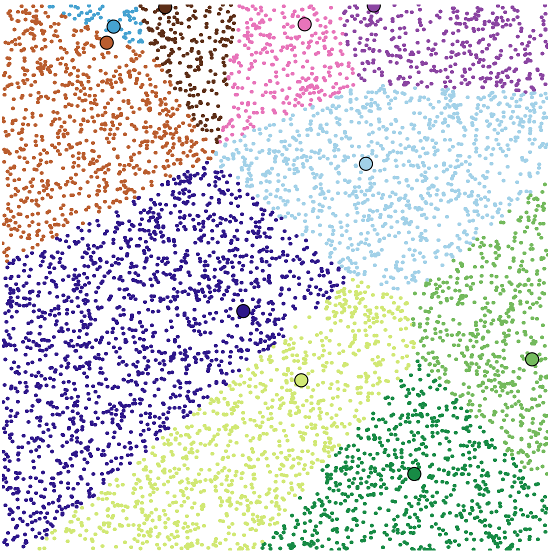
Goal of this tutorial part is to implement the *k*-means clustering algorithm.



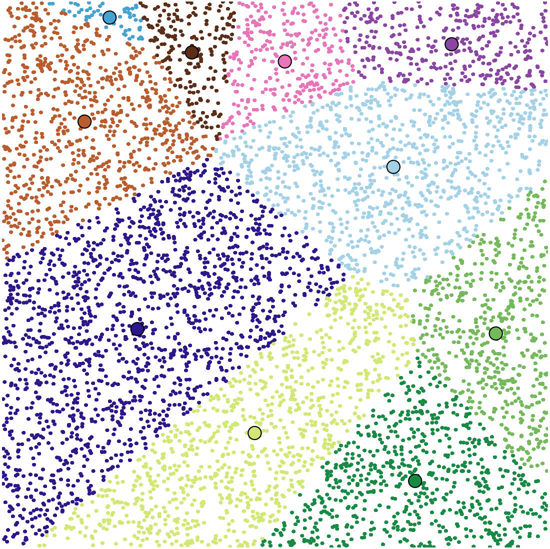
The algorithm works as follows:

1. Given are a set of d-dimensional points and a target number of clusters k.
2. Select k initial cluster center points at random.
3. Then attempt to improve the centers by iteratively calculating new centers. This is done by classifying all points and associating them with their nearest center, and then taking the mean of all points associated to one cluster as the new center.
4. This will be repeated a constant number of iterations.

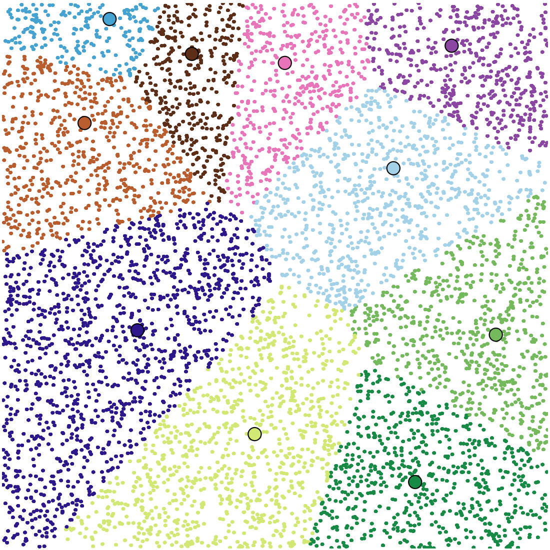
Tutorial: *k*-Means Iterations (pre 1)



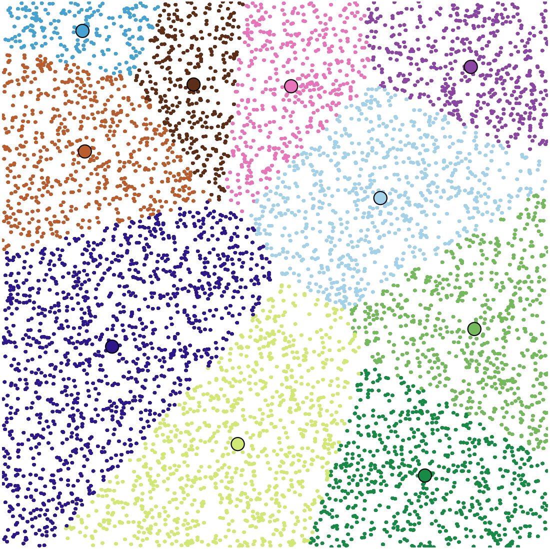
Tutorial: *k*-Means Iterations (post 1)



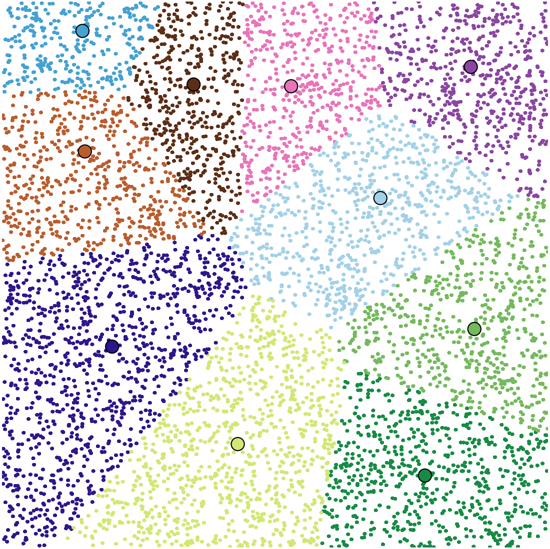
Tutorial: *k*-Means Iterations (pre 2)



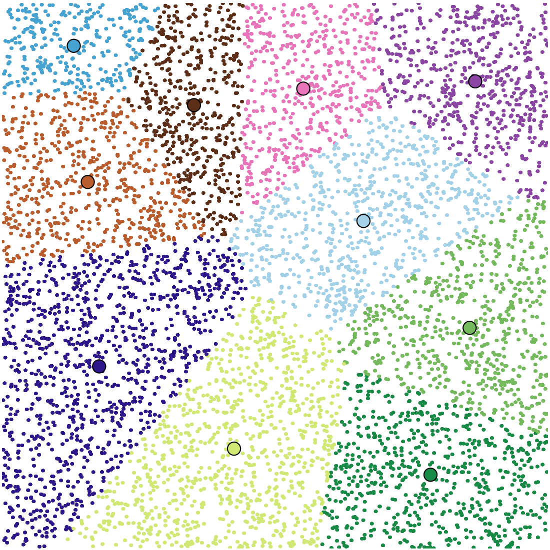
Tutorial: *k*-Means Iterations (post 2)



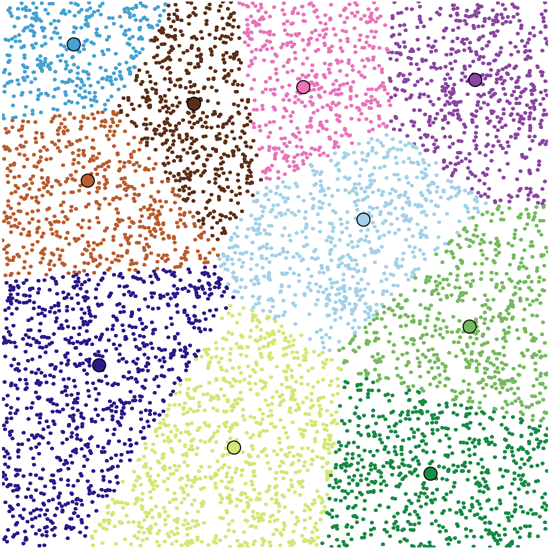
Tutorial: *k*-Means Iterations (pre 3)



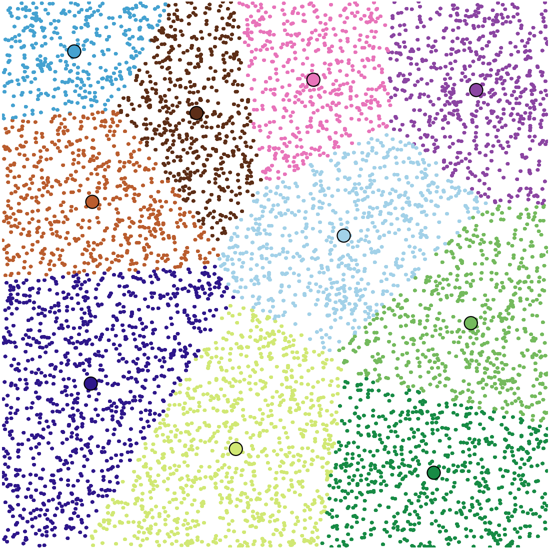
Tutorial: *k*-Means Iterations (post 3)



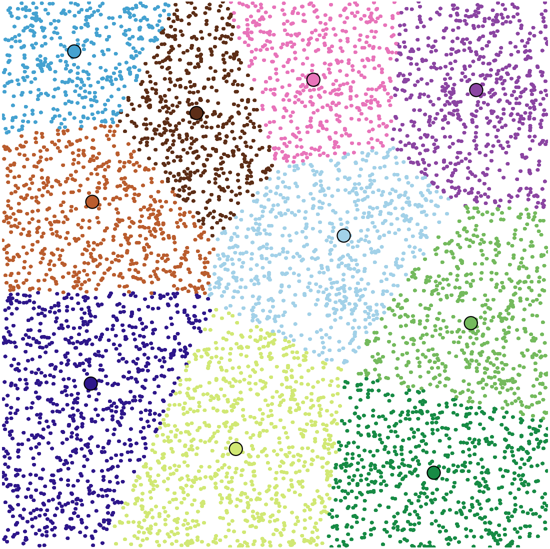
Tutorial: *k*-Means Iterations (pre 4)



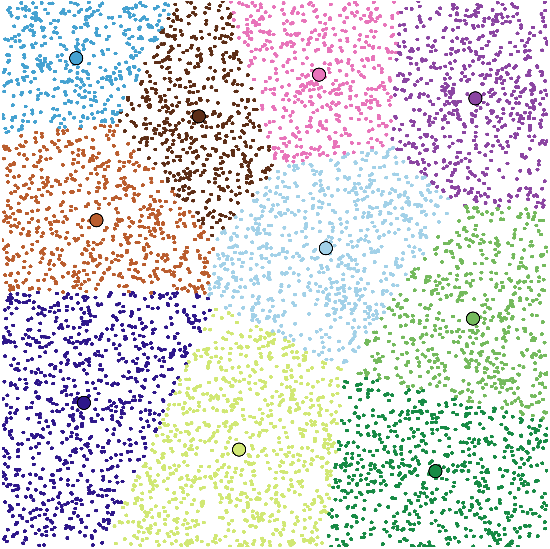
Tutorial: *k*-Means Iterations (post 4)



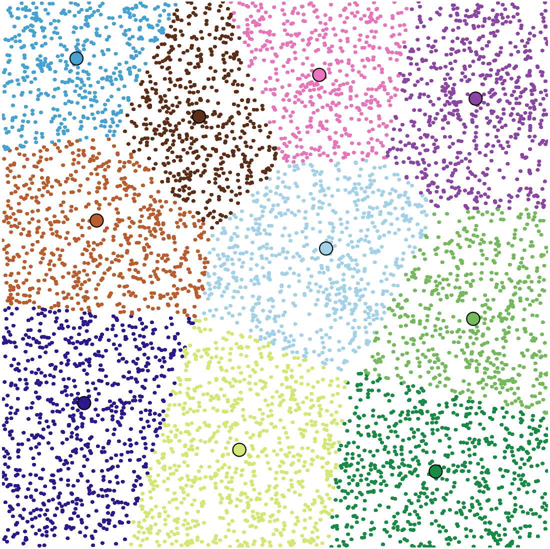
Tutorial: *k*-Means Iterations (pre 5)



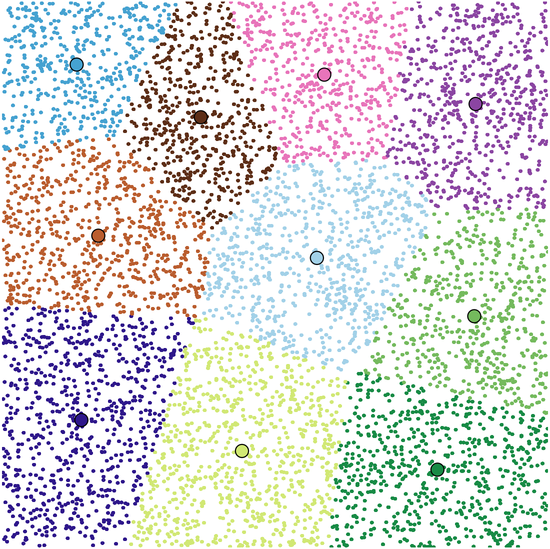
Tutorial: *k*-Means Iterations (post 5)



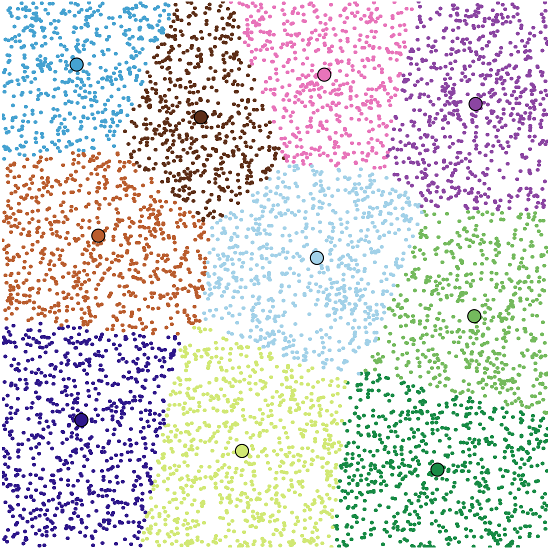
Tutorial: *k*-Means Iterations (pre 6)



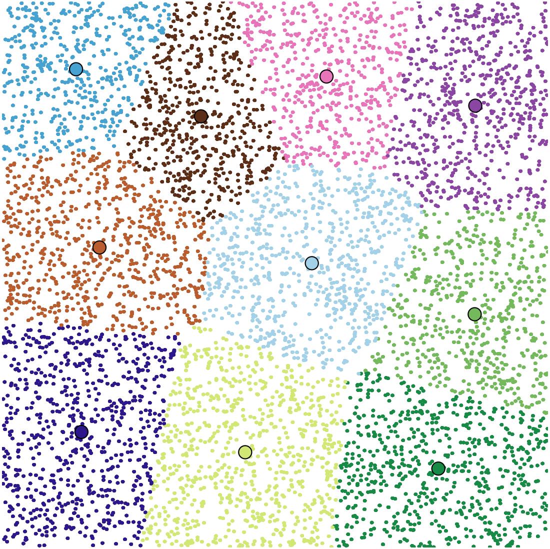
Tutorial: *k*-Means Iterations (post 6)



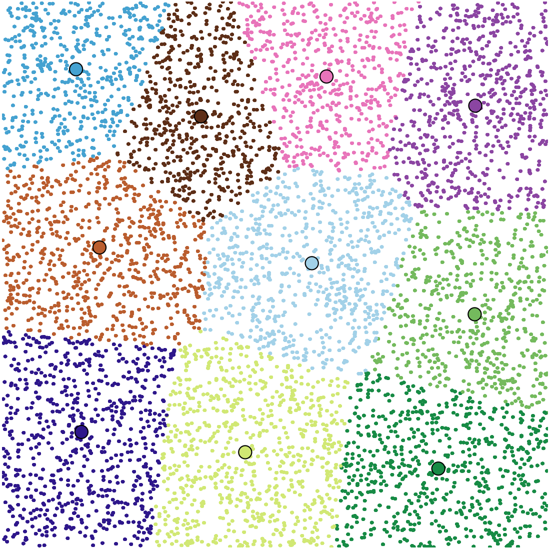
Tutorial: *k*-Means Iterations (pre 7)



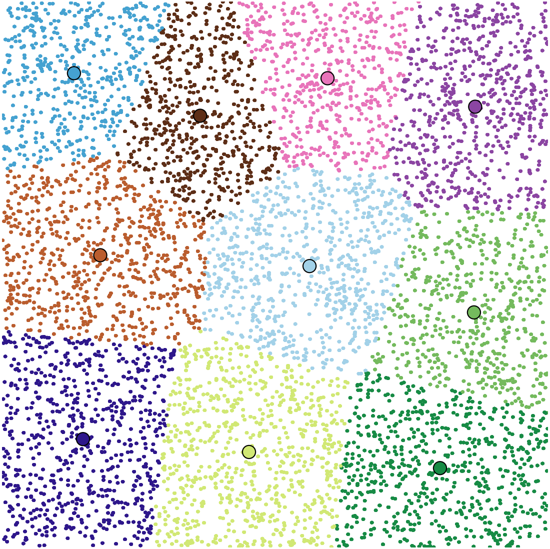
Tutorial: *k*-Means Iterations (post 7)



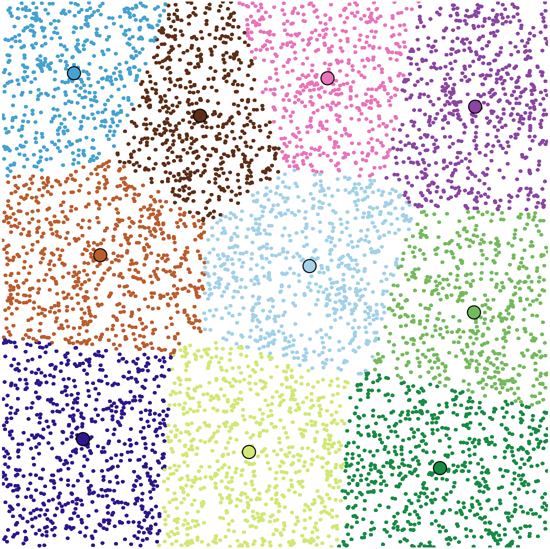
Tutorial: *k*-Means Iterations (pre 8)



Tutorial: *k*-Means Iterations (post 8)



Tutorial: *k*-Means Iterations (stop)



K-Means: Printable 2D-Points

Step 1: Make a 2D struct "Point" and generate random points.

- Use the following `Point` struct with `ostream` operator:

```
//! A 2-dimensional point with double precision
struct Point {
    //! point coordinates
    double x, y;
};
//! make ostream-able for Print()
std::ostream& operator << (std::ostream& os, const Point& p) {
    return os << '(' << p.x << ',' << p.y << ')';
}
```

- Use `Generate` to make random points, `Print`, and `Cache` them.
- Use the script `points2svg.py` to display the "`(x,y)`" lines.

K-Means: Map to Random Centers

Step 2: Map points to randomly selected centers.

- Use `Sample` to select random initial centers, `Print` them.
- `Map` each Point to its closest center.
- Maybe add a `distance` method to your Point and refactor.
- What should the `Map` output for the next step?
  What is the next step?

K-Means: Calculate Better Centers

Step 3: Calculate better centers by reducing all points.

- Next step is to use `ReduceByKey` or `ReduceToIndex` to calculate the mean of all points associated with a center.
- Key idea: make a second struct `PointTarget` containing Point and new target center id.
- Reduce all structs with same target center id and calculate the vector sum and the number of points associated.
- To do this, create a third struct `PointSumCount` containing Point, vector sum, and a counter.
- Maybe add add and scalar multiplication operators to Point.

K-Means: Iterate!

Step 4: Iterate the process 10 times.

- Collect the new centers on all hosts with `AllGather`.
- Add a `for` loop for iteration.

Bonus Step 5: Add input and output to/from text files.

Bonus Step 6: Instead of 10 iterations, calculate the distance that centers moved and break if below a threshold.

Bonus Step 7: Calculate the “error” of the centers, which is the total distance of all points to their cluster center.

Bonus Step 7: Run your program on the cluster with a large dataset.

3 The Thrill Framework

- Thrill's DIA Abstraction and List of Operations
- Tutorial: Playing with DIA Operations
- Execution of Collective Operations in Thrill
- Tutorial: Running Thrill on a Cluster
- Tutorial: Logging and Profiling
- Going Deeper into Thrill
- Tutorial: First Steps towards K-Means
- Conclusion

Thoughts on the Architecture

Thrill's Sweet Spot

- C++ toolkit for implementing distributed algorithms quickly.
- Platform to engineer and evaluate distributed primitives.
- Efficient processing of small items and pipelining of primitives.
- Platform for implementing on-the-fly compiled queries?

Open Questions

- Compile-time optimization only – no run-time algorithm selection or (statistical) knowledge about the data.
- Assumes *h* identical hosts constantly running, (the old MPI/HPC way, Hadoop/Spark do block-level scheduling).
- Memory management
- Predictability and scalability to 1 million cores

Current and Future Work

Case Studies:

- Five suffix sorting algorithms [B, Gog, Kurpicz, BigData'18]
- Louvain graph clustering [Hamann et al. Euro-Par'18]
- Process scientific data on HPC (poster) [Karabin et al. SC'18]
- More: stochastic gradient descent, triangle counting, etc.

Ideas for Future Work:

- Distributed rank()/select() and wavelet tree construction.
- Beyond `DIA<T>`? `Graph<V,E>`? `DenseMatrix<T>`?
- Fault tolerance and communication efficient scalability.

Thank you for your attention!