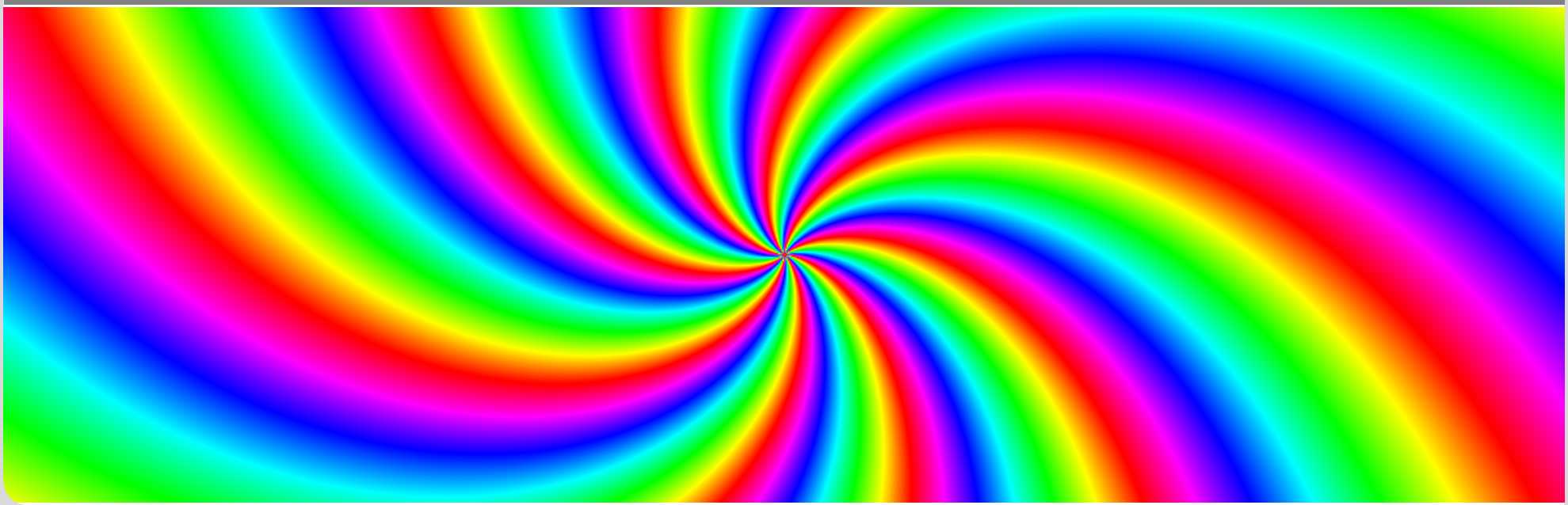
Karlsruhe Institute of Technology

Thrill : High-Performance Algorithmic Distributed Batch Data Processing with C++

Timo Bingmann, Michael Axtmann, Peter Sanders, Sebastian Schlag, and 6 Students | 2017-09-01

INSTITUTE OF THEORETICAL INFORMATICS – ALGORITHMICS



Weak-Scaling Benchmarks

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

- Reduce text files from CommonCrawl web corpus.

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

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

TeraSort – $h \cdot 16\,\text{GiB}$

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

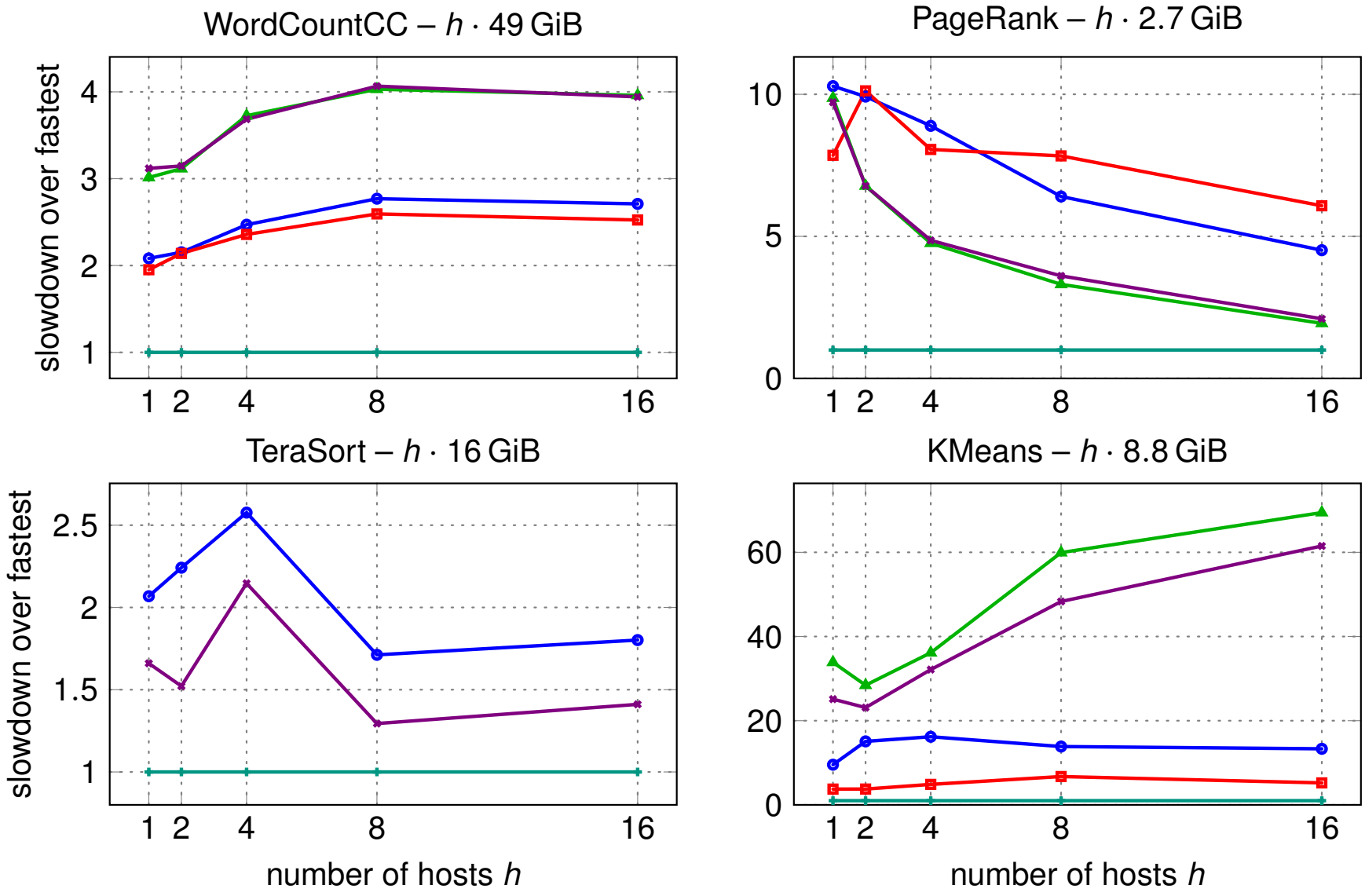
K-Means – $h \cdot 8.8\,\text{GiB}$

- 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



Example T = [dbadcbccbabdcc$]

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

SA_i	LCP_i	$T_{SA_i \ldots n}$
14	-	$
9	0	a b d c c $
2	1	a d c b c c b a b d c c $
8	0	b a b d c c $
1	2	b a d c b c c b a b d c c $
5	1	b c c b a b d c c $
10	1	b d c c $
13	0	c $
7	1	c b a b d c c $
4	2	c b c c b a b d c c $
12	1	c c $
6	2	c c b a b d c c $
0	0	d b a d c b c c b a b d c c $
3	1	d c b c c b a b d c c $
11	2	d c c $

Suffix Sorting with DC3: Example

	0	1	2	3	4	5	6	7	8	9	10
$T =$	d	b	a	c	b	a	c	b	d	$	$

$= [\, t_i \,]_{i=0,\ldots,n-1}$

triples	(bac,1),	(bac,4),	(bd$,7),	(acb,2)	(acb,5),	(d$$,8)
sorted	(acb,2)	(acb,5),	(bac,1),	(bac,4),	(bd$,7),	(d$$,8)
equal 0/1	0	0	1	0	1	1
prefix sum	0	0	1	1	2	3

$R =$ 1 1 2 0 0 3 $

r_1 r_4 r_7 r_2 r_5 r_8

$\mathrm{SA}_R =$ 3 4 0 1 2 5 $ $\mathrm{ISA}_R =$ 2 3 4 0 1 5 $

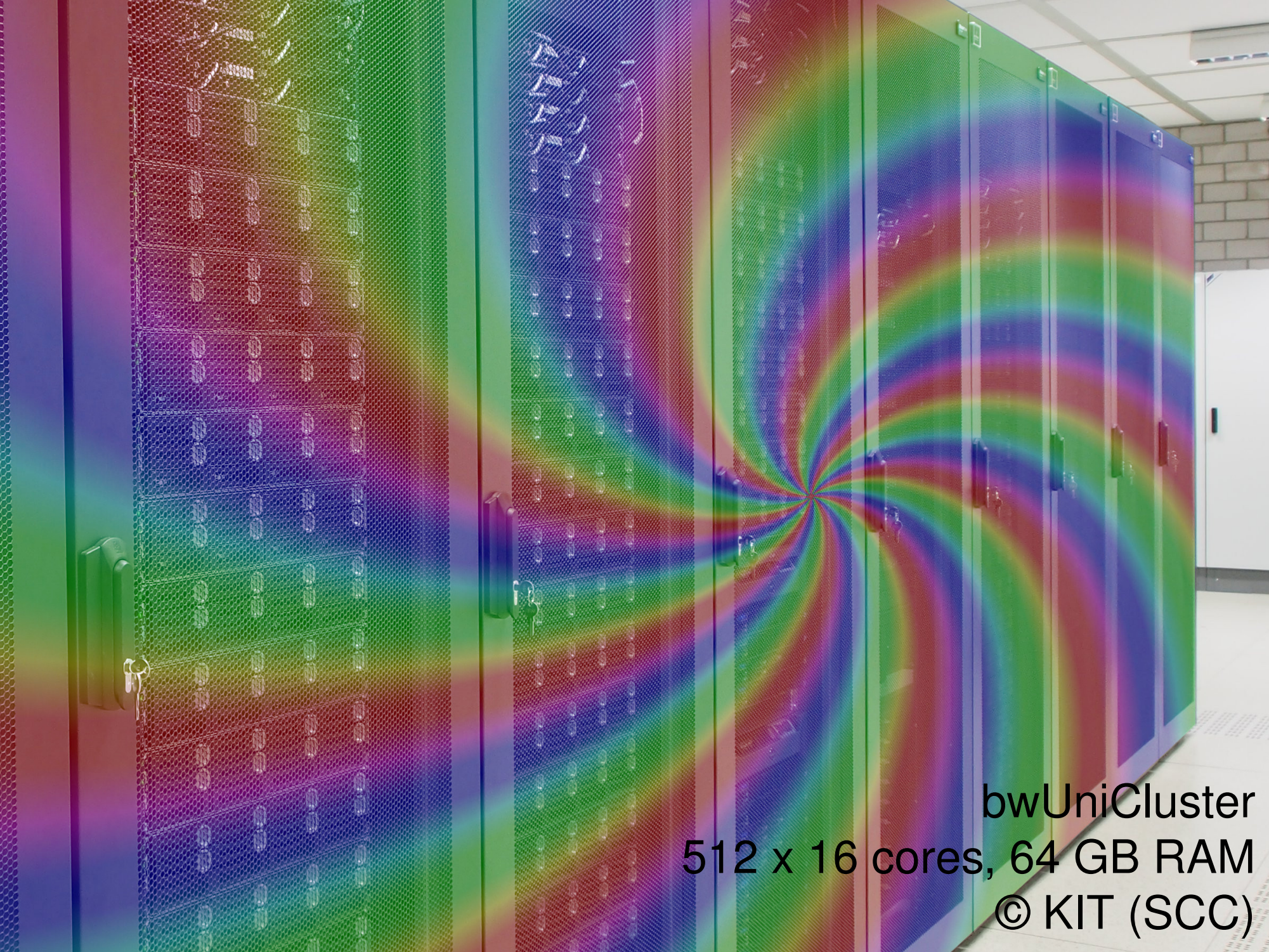
$S_0 = [(\mathtt{d},\mathtt{b},2,0,0),(\mathtt{c},\mathtt{b},3,1,3),(\mathtt{c},\mathtt{b},4,5,6)]$ $(t_i, t_{i+1}, r_{i+1}, r_{i+2}, i)$

$S_1 = [(2,\mathtt{b},0,1),(3,\mathtt{b},1,4),(4,\mathtt{b},5,7)]$ $(r_{i+1}, t_{i+1}, r_{i+2}, i+1)$

$S_2 = [(0,\mathtt{a},\mathtt{c},3,2),(1,\mathtt{a},\mathtt{c},4,5),(5,\mathtt{d},\$,6,8)]$ $(r_{i+2}, t_{i+2}, t'_{i+3}, r'_{i+4}, i+2)$

$\mathrm{SA}_T = \mathrm{Merge}(\mathrm{Sort}(S_0), \mathrm{Sort}(S_1), \mathrm{Sort}(S_2))$

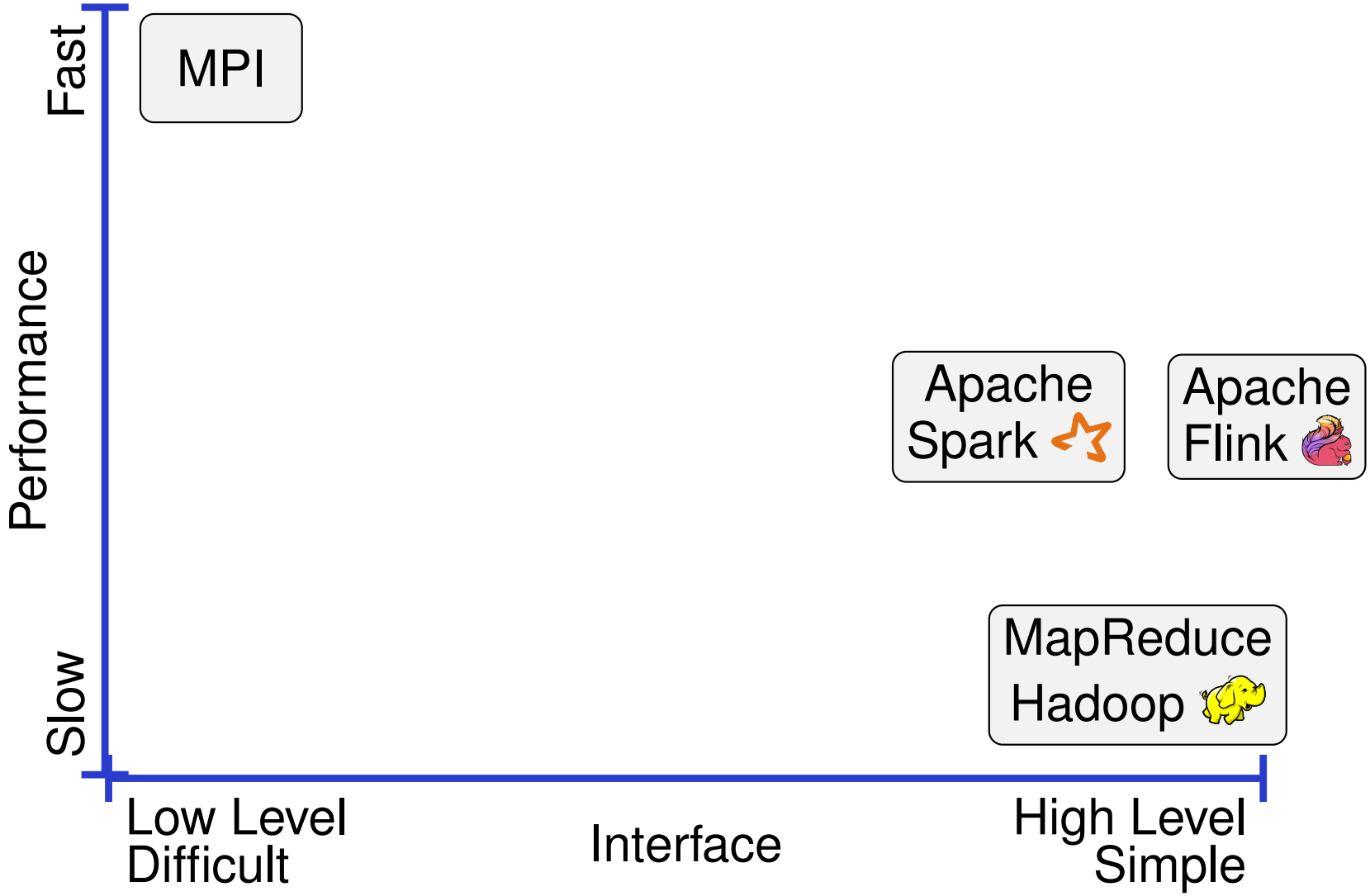
$\Theta(\mathrm{sort}(n))$


bwUniCluster
512 x 16 cores, 64 GB RAM
© KIT (SCC)

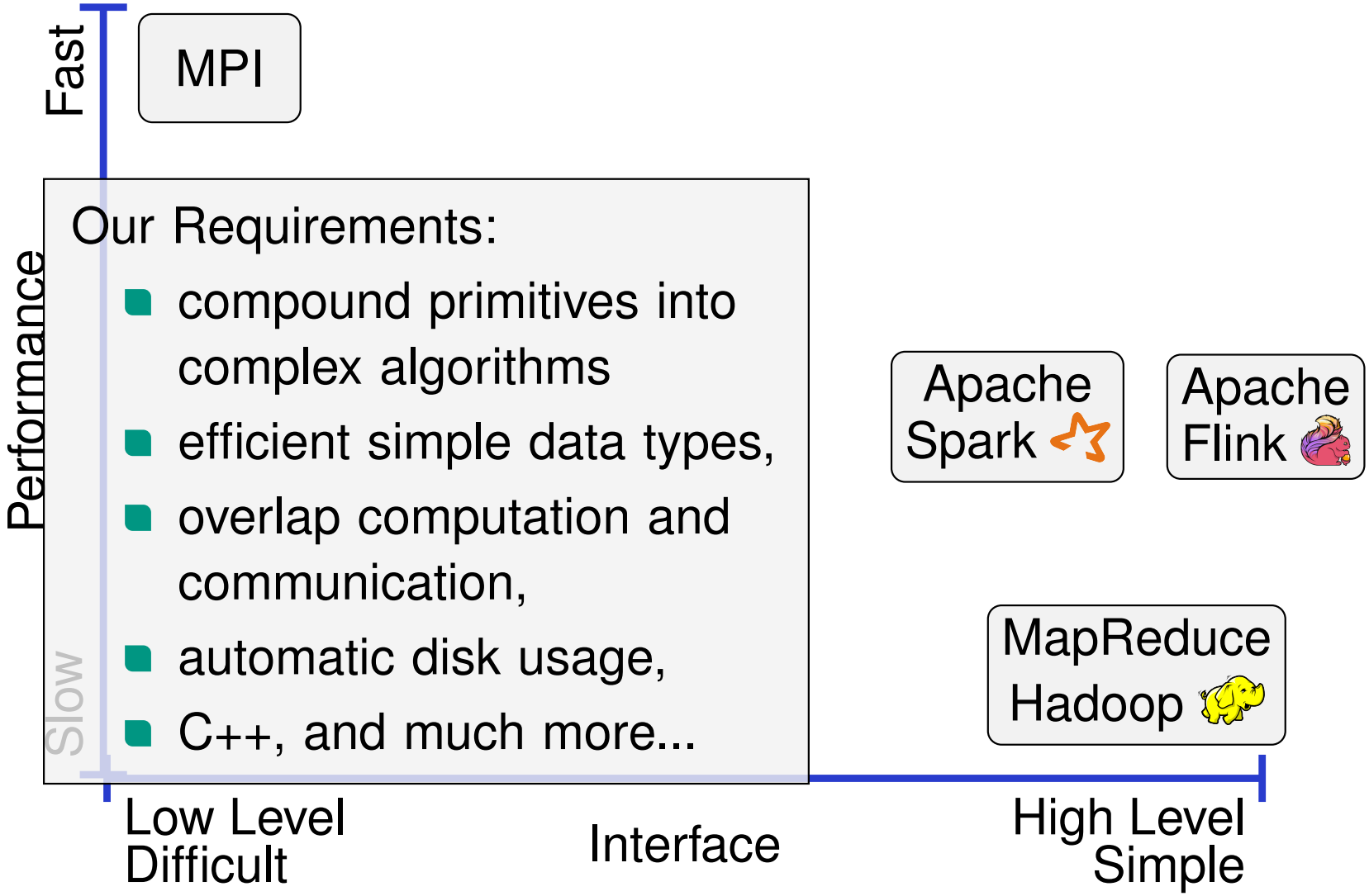
Flavours of Big Data Frameworks

- Batch Processing
  Google's MapReduce, Hadoop MapReduce  , Apache Spark  , Apache Flink  (Stratosphere), Google's FlumeJava.
- High Performance Computing (Supercomputers)
  MPI
- Real-time Stream Processing
  Apache Storm  , Apache Spark Streaming, Google's MillWheel.
- Interactive Cached Queries
  Google's Dremel, Powerdrill and BigQuery, Apache Drill  .
- Sharded (NoSQL) Databases and Data Warehouses
  MongoDB  , Apache Cassandra, Apache Hive, Google BigTable, Hypertable, Amazon RedShift, FoundationDB.
- Graph Processing
  Google's Pregel, GraphLab  , Giraph  , GraphChi.
- Time-based Distributed Processing
  Microsoft's Dryad, Microsoft's Naiad.

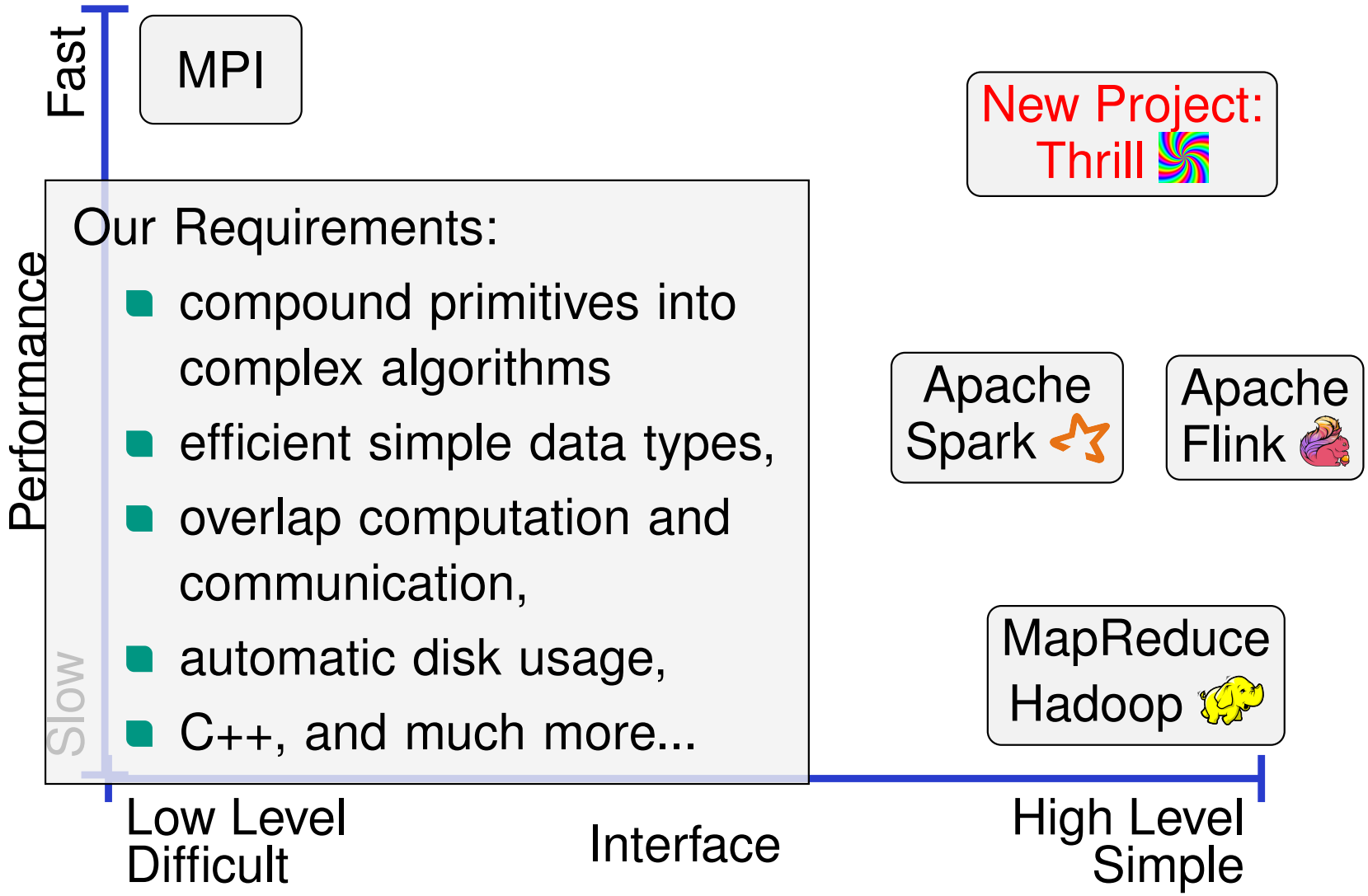
Big Data Batch Processing



Big Data Batch Processing



Big Data Batch Processing



Karlsruher Institut für Technologie

Projektpraktikum: Verteilte Datenverarbeitung mit MapReduce

Timo Bingmann, Peter Sanders und Sebastian Schlag | 21. Oktober 2014 @ PdF Vorstellung

INSTITUTE OF THEORETICAL INFORMATICS – ALGORITHMICS

Thrill's Design Goals

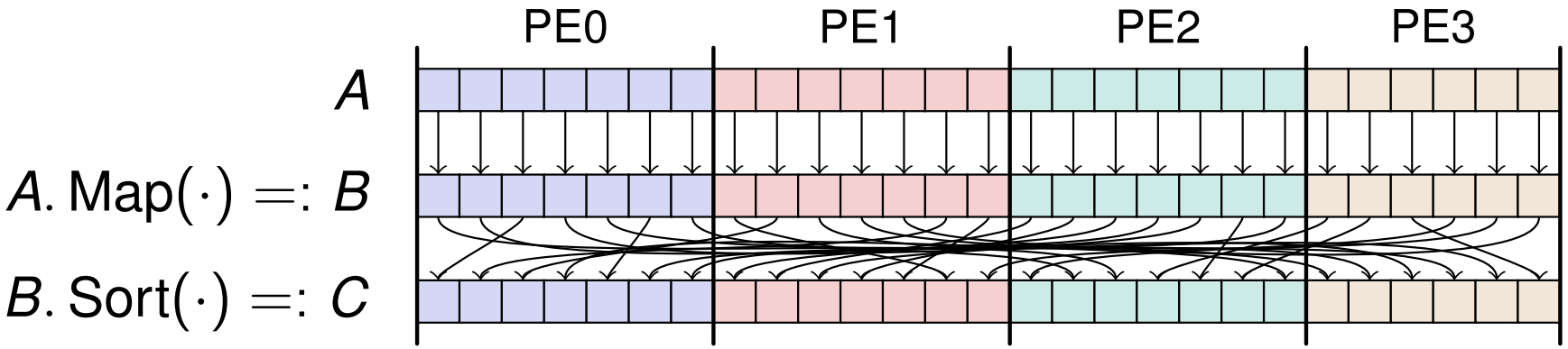
- An easy way to program distributed algorithms in C++.
- Distributed arrays of small items (characters or integers).
- High-performance, parallelized C++ operations.
- Locality-aware, in-memory computation.
- Transparently use disk if needed
  $\Rightarrow$ external memory or cache-oblivious algorithms.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

Current Status:

- open-source prototype at http://github.com/thrill/thrill.

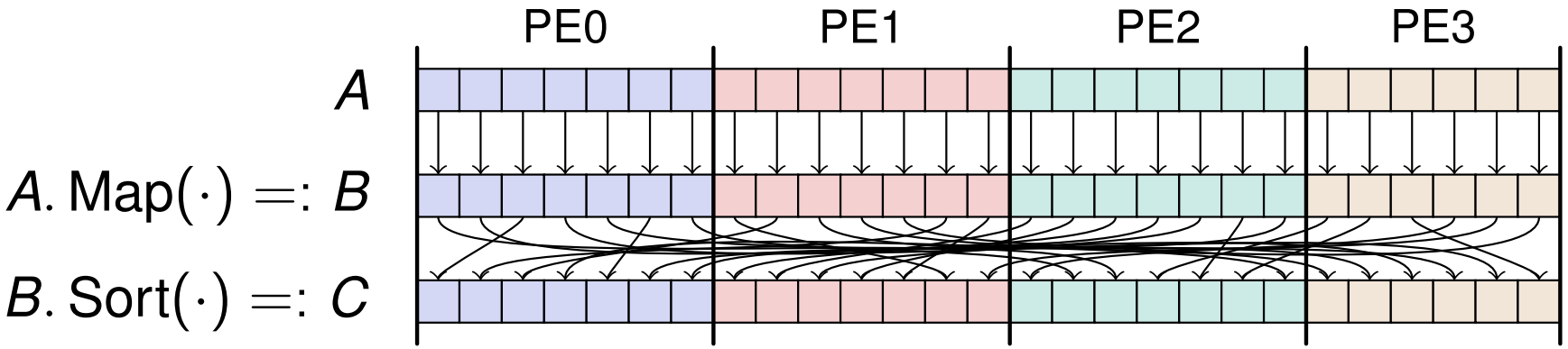
Distributed Immutable Array (DIA)

- 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)

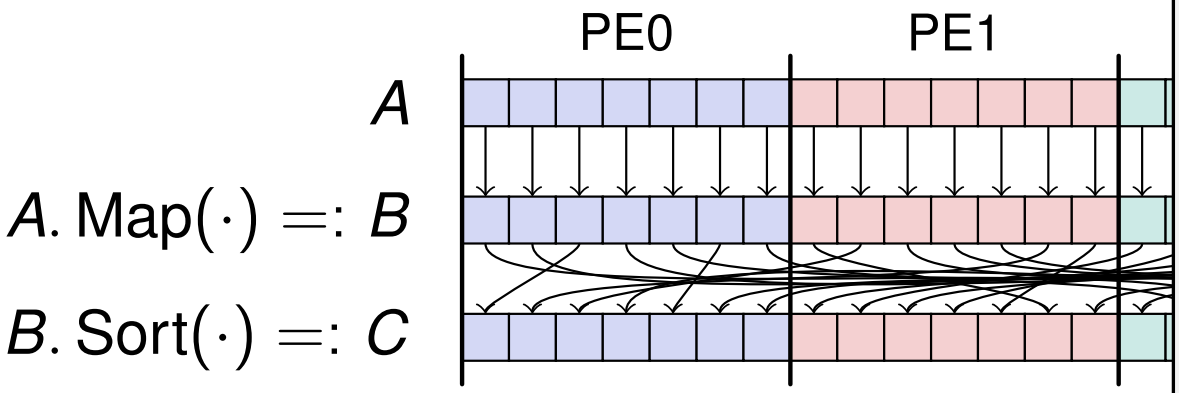
- User Programmer's View:
  - `DIA<T>` = result of an operation (local or distributed).
  - Model: distributed array of items T on the cluster
  - Cannot access items directly, instead use transformations and actions.

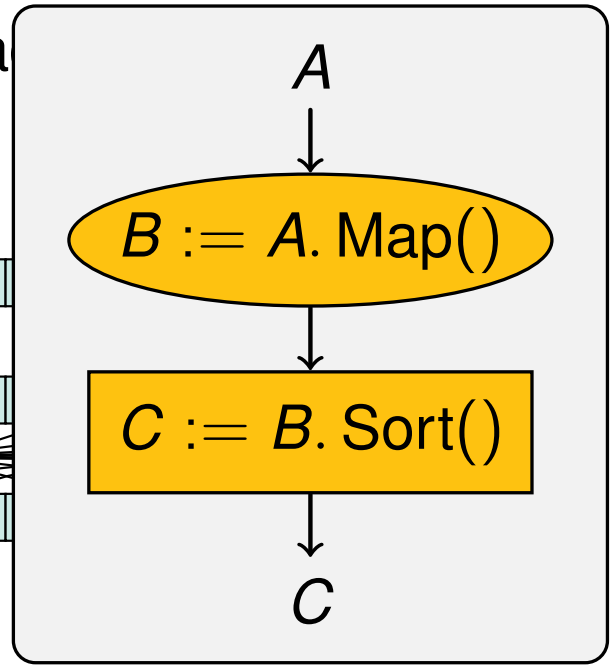


- Framework Designer's View:
  - Goals: distribute work, optimize execution on cluster, add redundancy where applicable. $\Longrightarrow$ build data-flow graph.
  - `DIA<T>` = chain of computation items
  - Let distributed operations choose "materialization".

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, instea
    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.

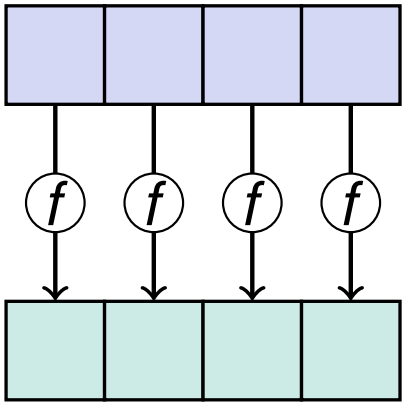
Sort()	Sort a DIA using comparisons.
ReduceBy()	Shuffle with Key Extractor, Hasher, and associative Reducer.
GroupBy()	Like ReduceBy, but with a general Reducer.
PrefixSum()	Compute (generalized) prefix sum on DIA.
Window$_k$()	Scan all k consecutive DIA items.
Zip()	Combine equal sized DIAs item-wise.
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)

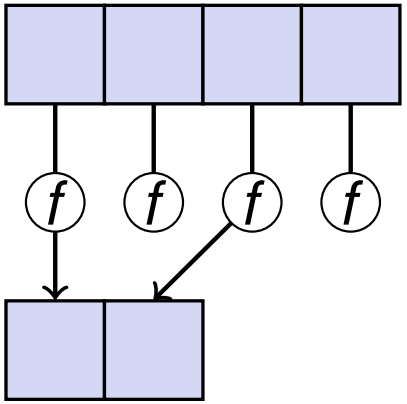
Map$(f) : \langle A \rangle \to \langle B \rangle$

$f : A \to B$



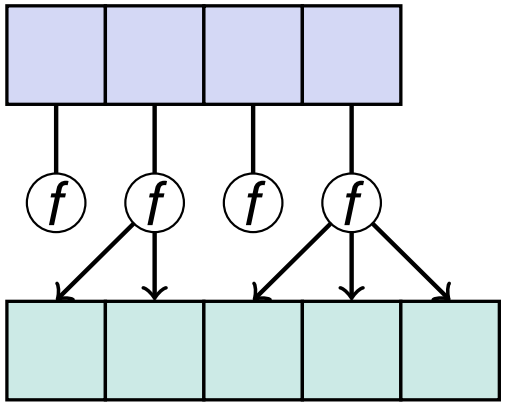
Filter$(f) : \langle A \rangle \to \langle A \rangle$

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



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

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



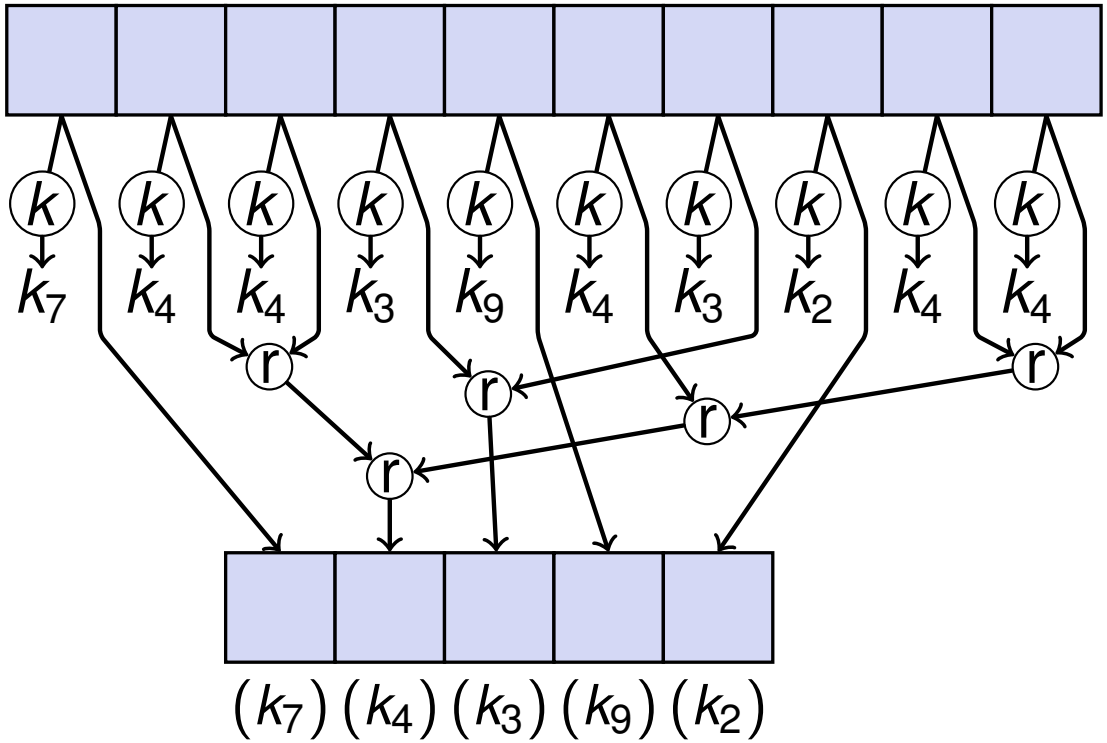
Currently: no rebalancing during LOps.

DOps: ReduceByKey

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

$k : A \to K$ key extractor

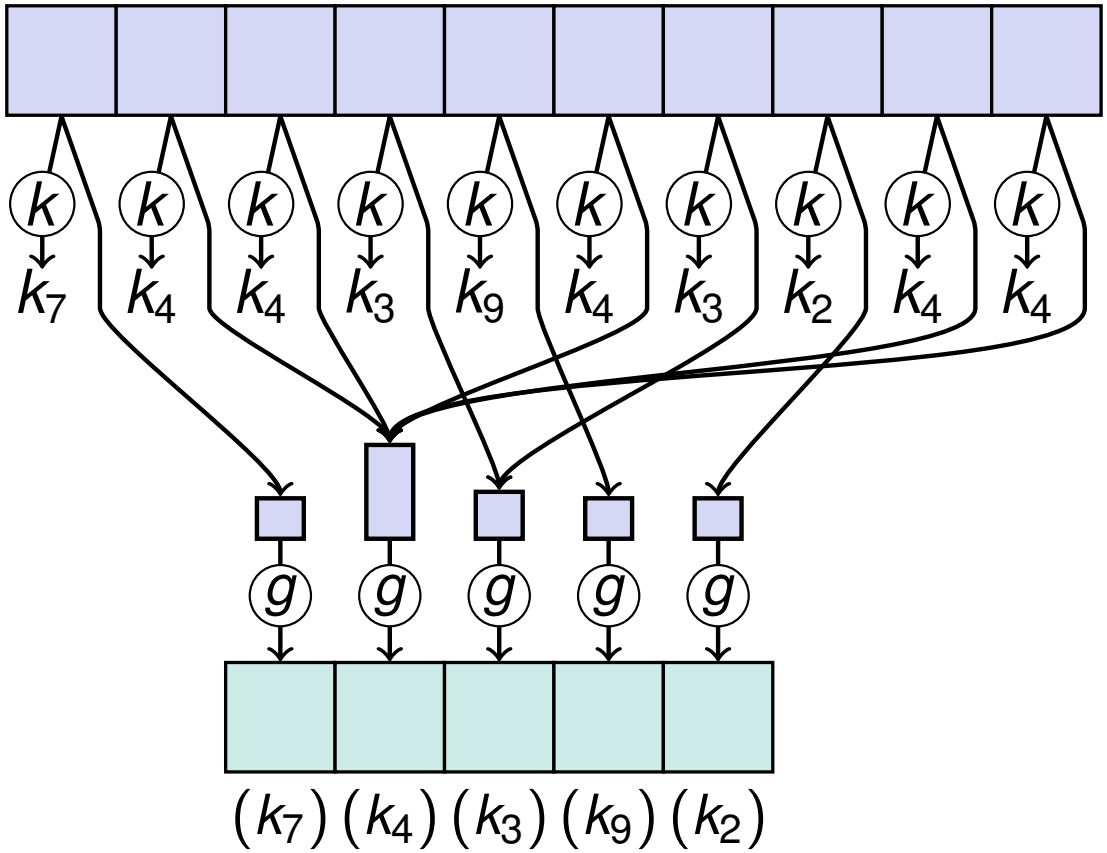
$r : A \times A \to A$ reduction



DOps: GroupByKey

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

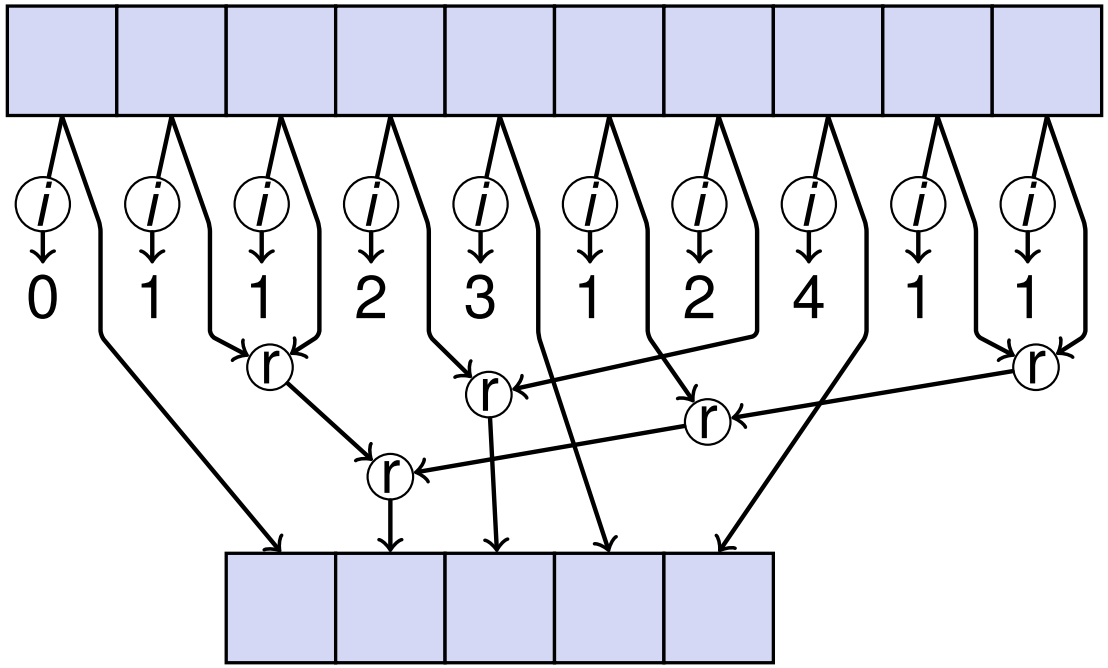
$k : A \rightarrow K$	key extractor
$g : \mathit{iterable}(A) \rightarrow B$	group function



DOps: ReduceToIndex

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

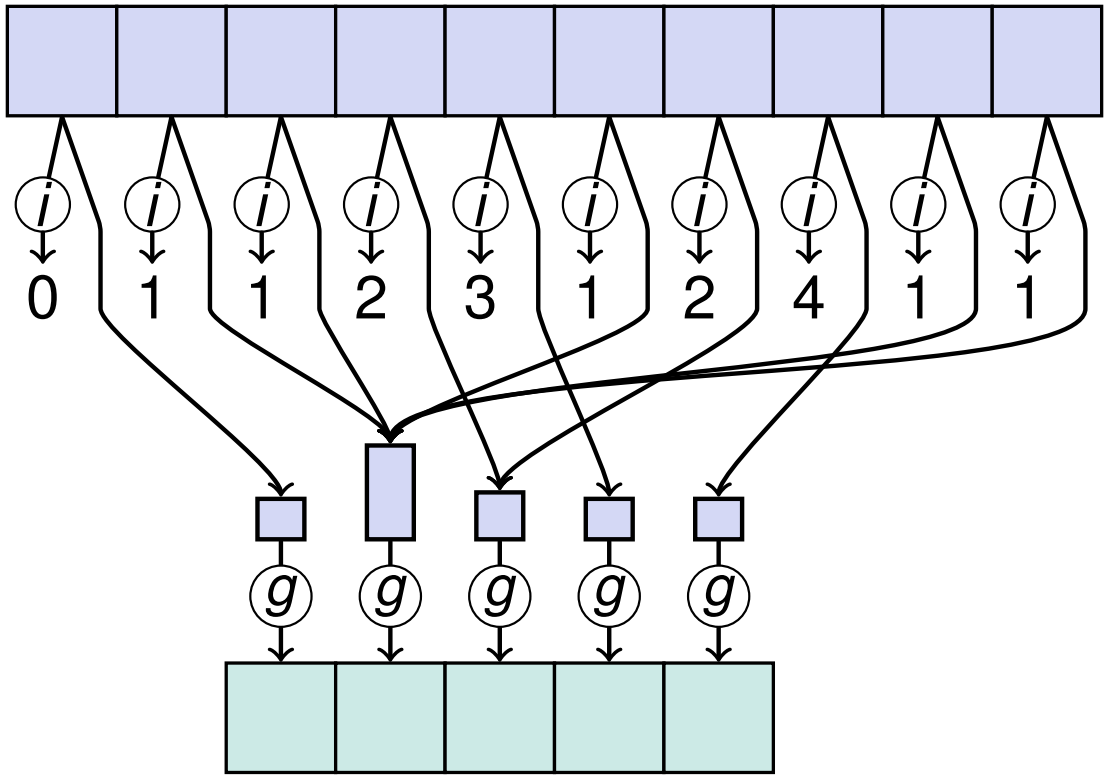
$i : A \to \{0..n-1\}$	index extractor
$n \in \mathbb{N}_0$	result size
$r : A \times A \to A$	reduction



DOps: GroupToIndex

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

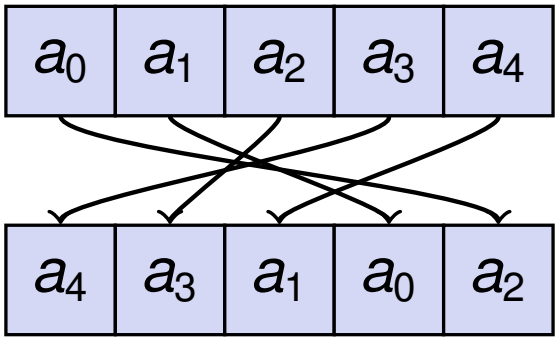


DOps: Sort and Merge

Sort$(o) : \langle A \rangle \rightarrow \langle A \rangle$

$o : A \times A \rightarrow \{false, true\}$

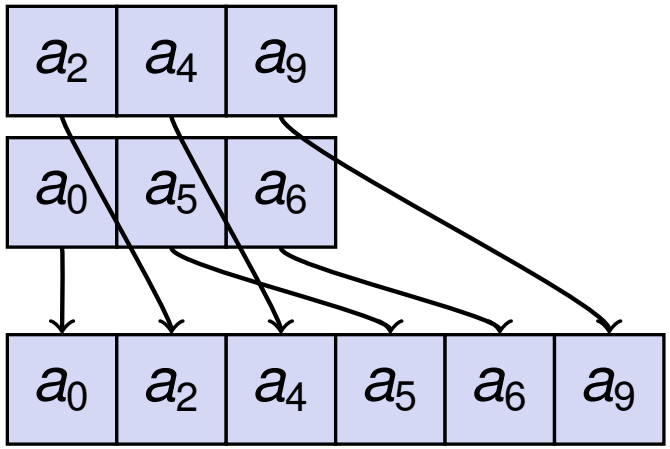
(less) order relation



Merge$(o) : \langle A \rangle \times \langle A \rangle \cdots \rightarrow \langle A \rangle$

$o : A \times A \rightarrow \{false, true\}$

(less) order relation

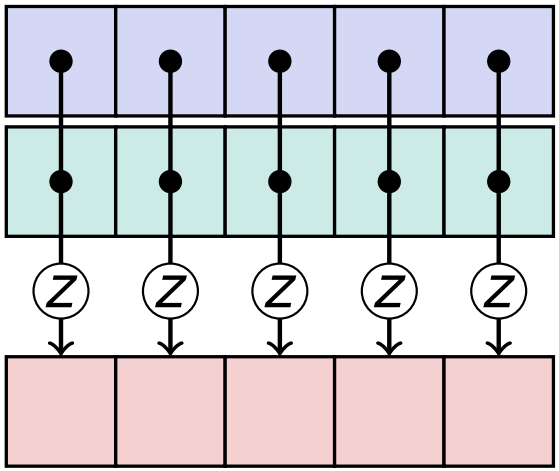


DOps: Zip and Window

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

$z : A \times B \rightarrow C$

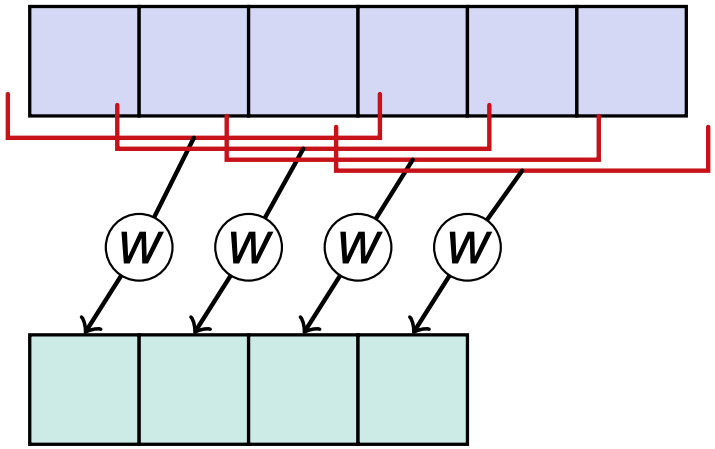
zip function



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

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

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

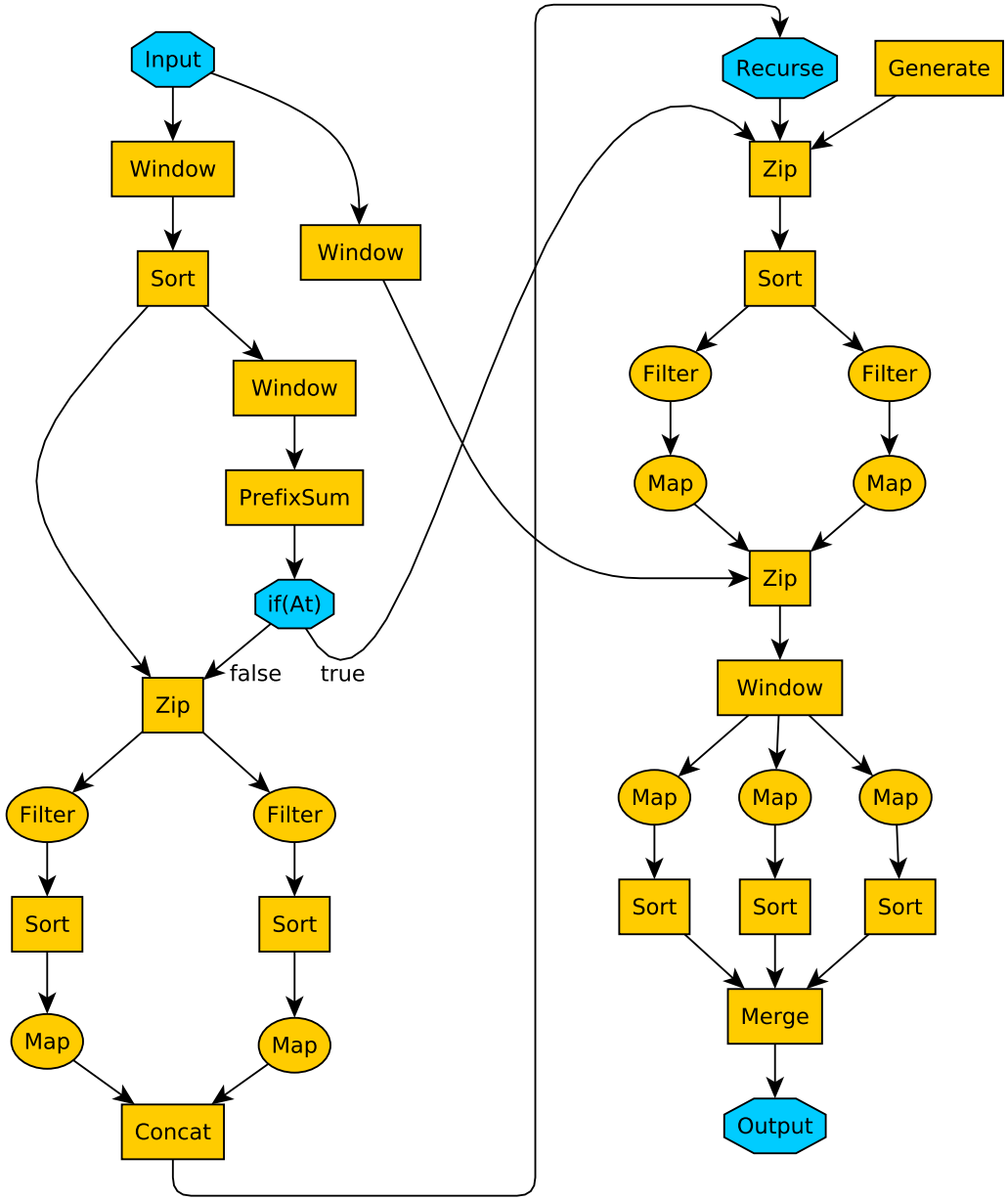


Example: WordCount in Thrill

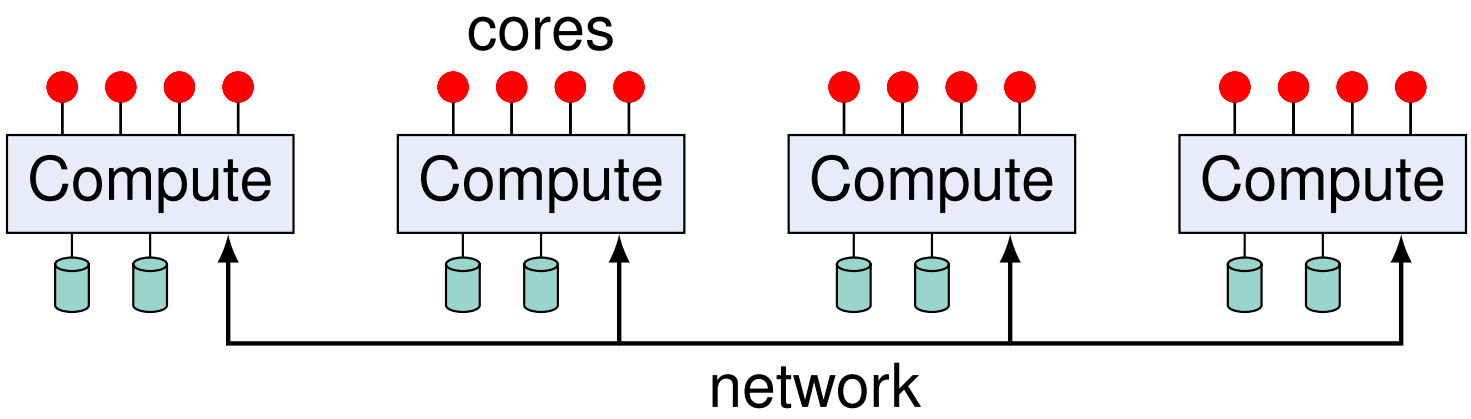
Karlsruhe Institute of Technology

```
1  using Pair = std::pair<std::string, size_t>;
2  void WordCount(Context& ctx, std::string input, std::string output) {
3      auto word_pairs = ReadLines(ctx, input)     // DIA<std::string>
4      .FlatMap<Pair>(
5          // flatmap lambda: split and emit each word
6          [](const std::string& line, auto emit) {
7              Split(line, ' ', [&](std::string_view sv) {
8                  emit(Pair(sv.to_string(), 1)); });
9      });                                         // DIA<Pair>
10     word_pairs.ReduceByKey(
11         // key extractor: the word string
12         [](const Pair& p) { return p.first; },
13         // commutative reduction: add counters
14         [](const Pair& a, const Pair& b) {
15             return Pair(a.first, a.second + b.second);
16     })                                          // DIA<Pair>
17     .Map([](const Pair& p) {
18         return p.first + ": " + std::to_string(p.second); })
19     .WriteLines(output);                        // DIA<std::string>
20 }
```

A Suffix Sorting Algorithm: DC3



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.

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

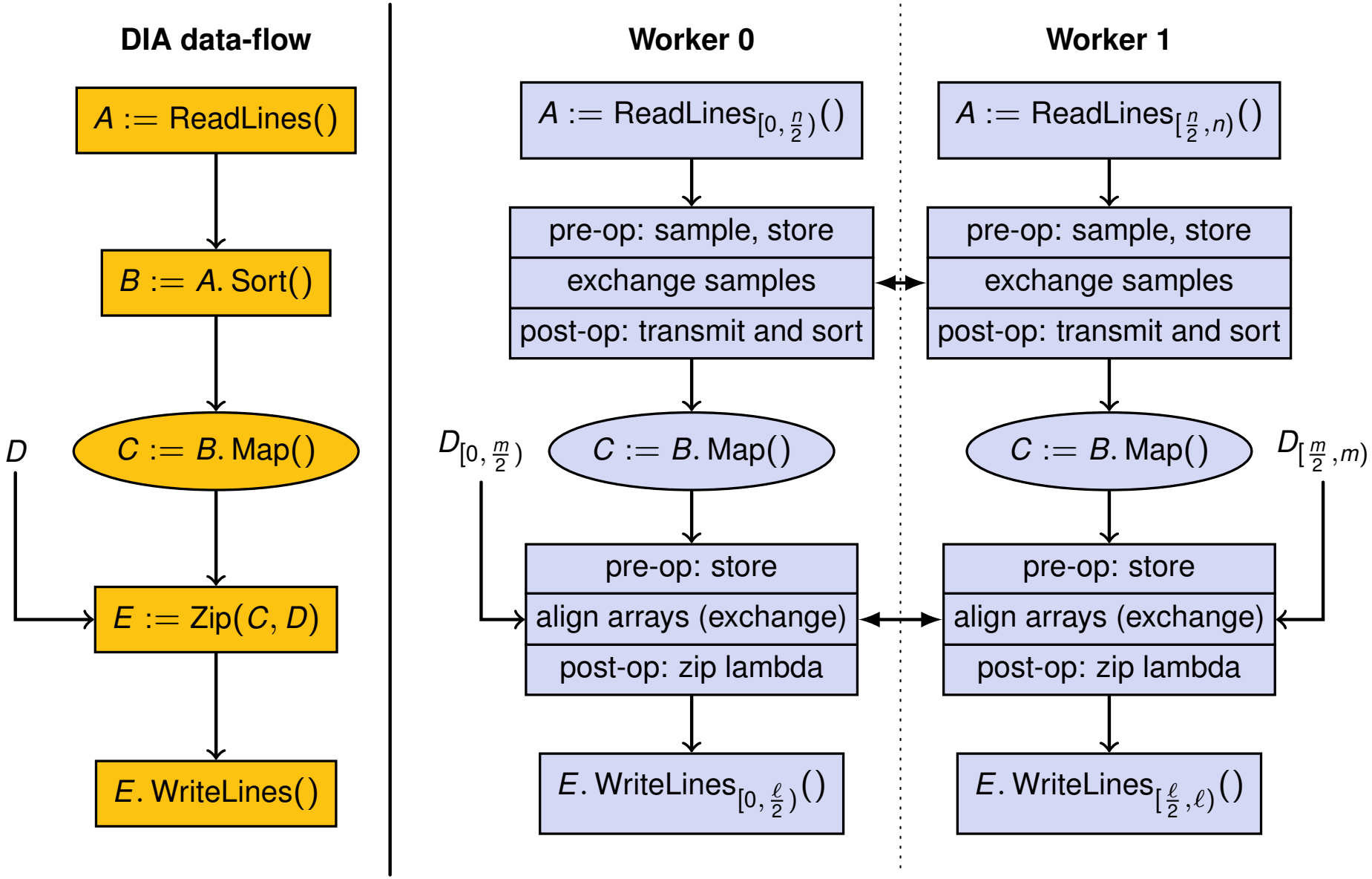
net: Network Layer
(Binomial Tree) Broadcast, Reduce, AllReduce, Async-Send/Recv, Dispatcher
Backends:

mock	tcp	mpi

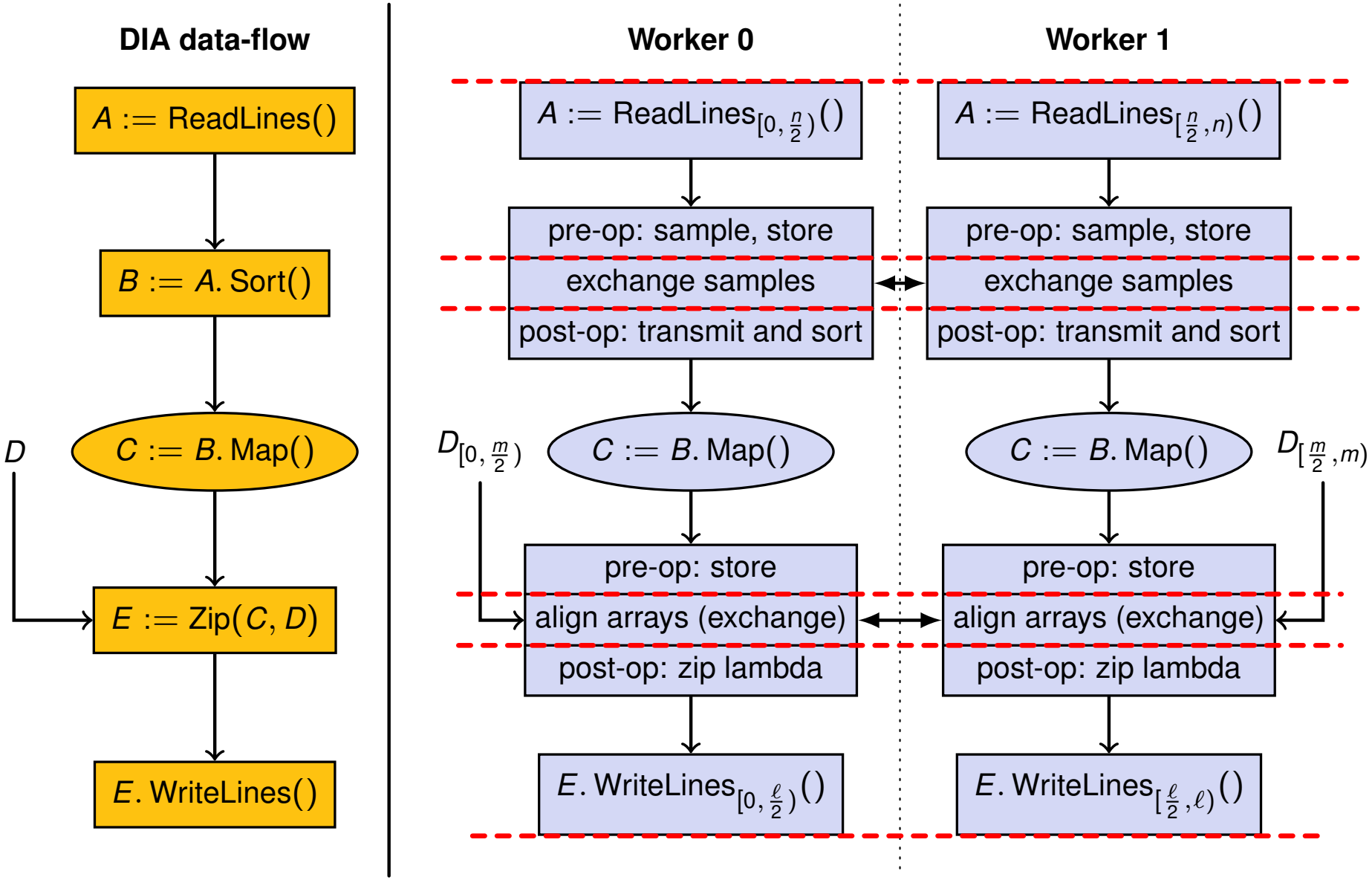
common: Common Tools
Logger, Delegates, Math, ...

mem: Memory Limitation
Allocators, Counting

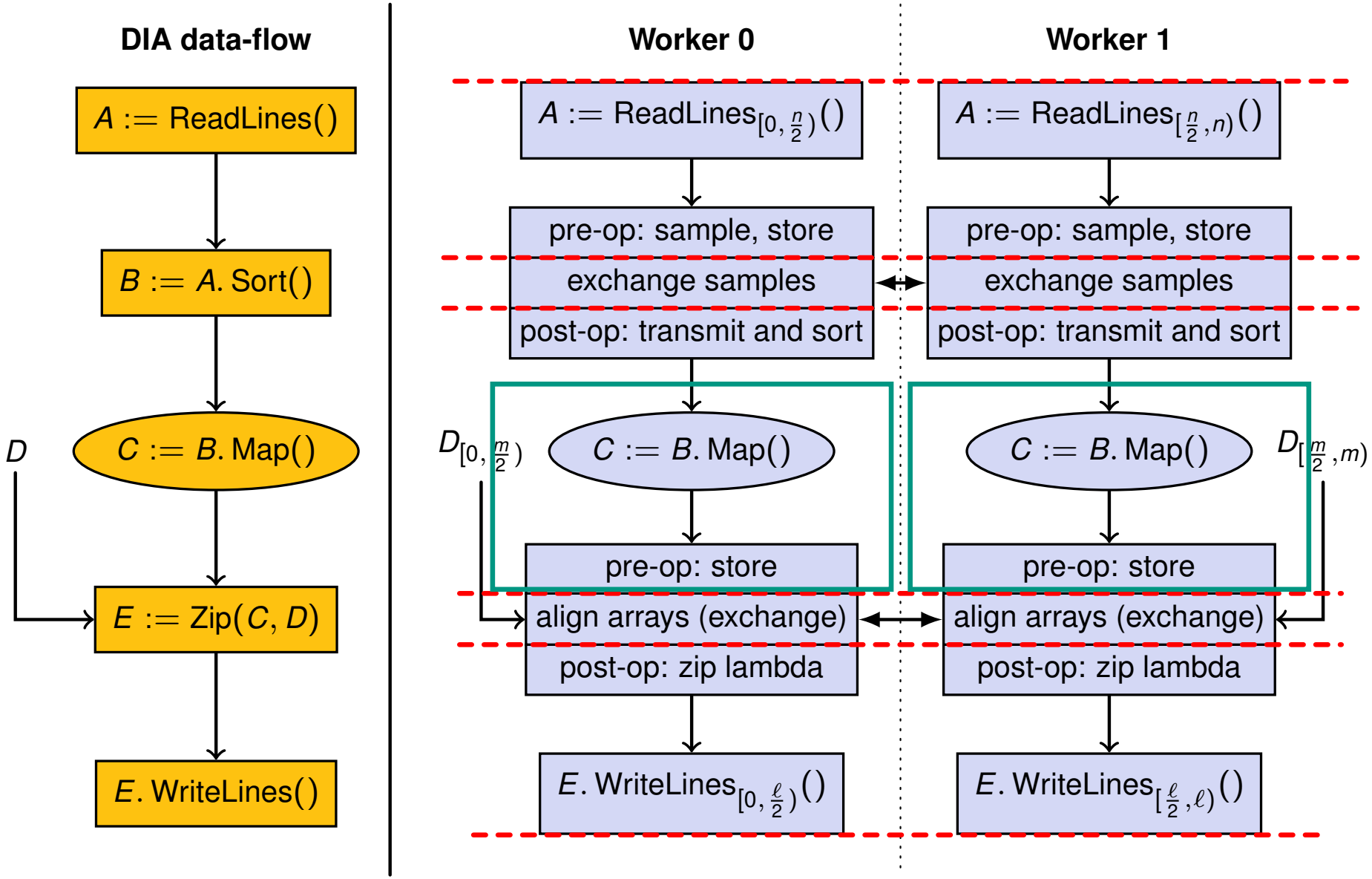
Mapping Data-Flow Nodes to Cluster



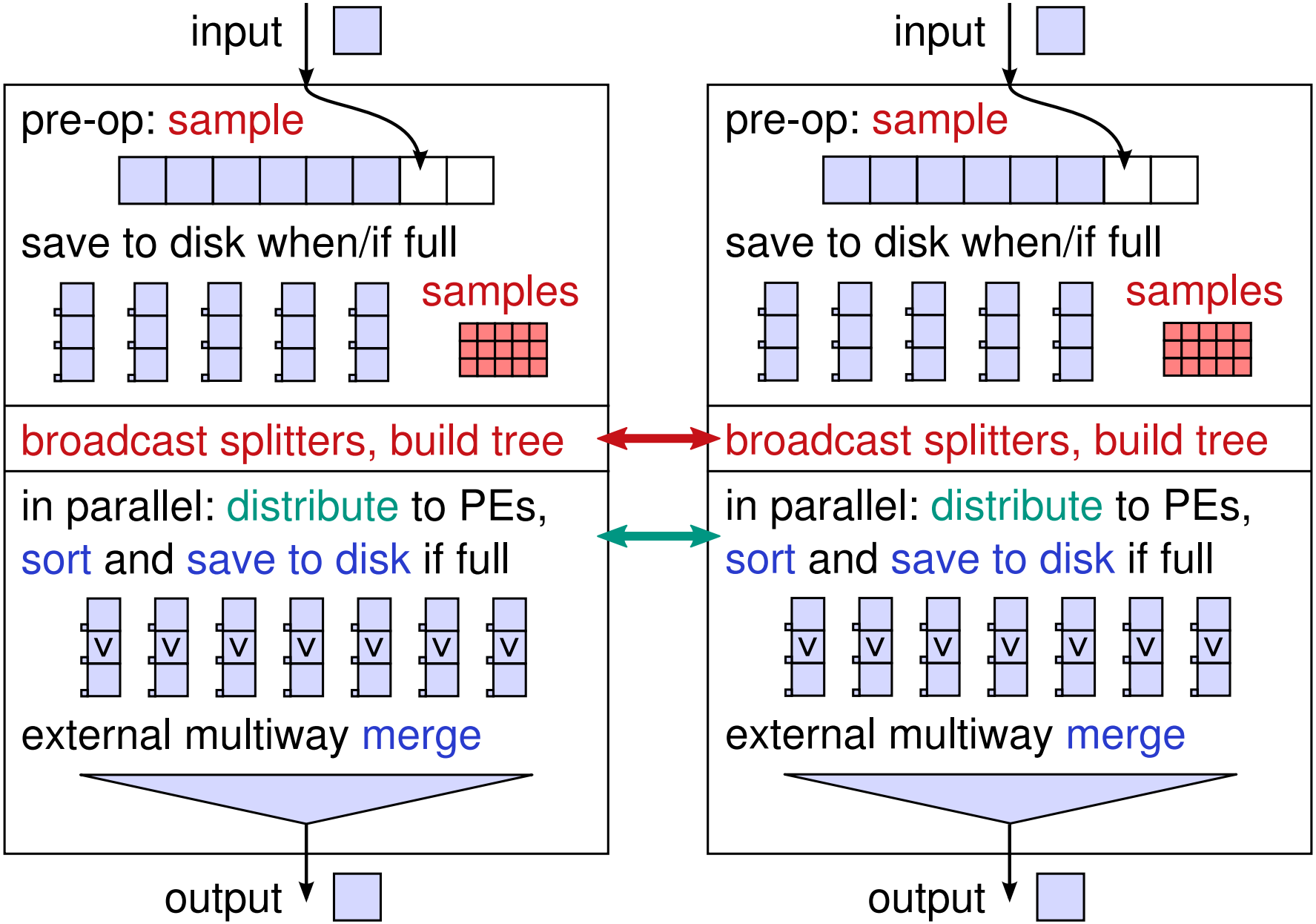
Mapping Data-Flow Nodes to Cluster



Mapping Data-Flow Nodes to Cluster



Sorting DOp


input
pre-op: sample
save to disk when/if full
samples
broadcast splitters, build tree
in parallel: distribute to PEs,
sort and save to disk if full
external multiway merge
output
input
pre-op: sample
save to disk when/if full
samples
broadcast splitters, build tree
in parallel: distribute to PEs,
sort and save to disk if full
external multiway merge
output

Weak-Scaling Benchmarks

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

- Reduce text files from CommonCrawl web corpus.

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

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

TeraSort – $h \cdot 16\,\text{GiB}$

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

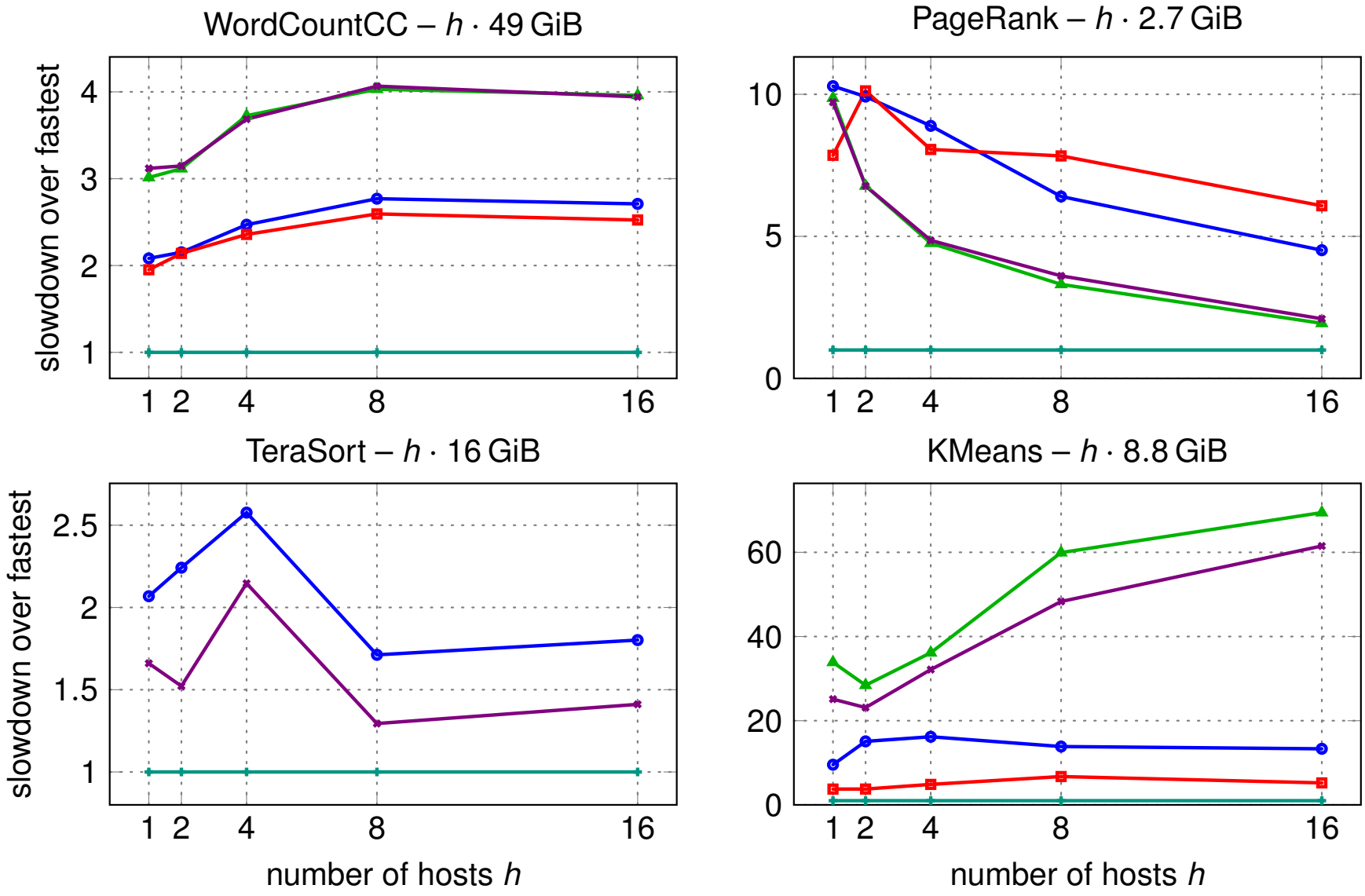
K-Means – $h \cdot 8.8\,\text{GiB}$

- 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



Thoughts on the Architecture

Thrill's Sweet Spot

- C++ toolkit for experimenting with distributed algorithms.
- 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 hosts constantly running, (the old MPI/HPC way of things, Hadoop/Spark do block-level scheduling).
- Memory management

K-Means Tutorial

Thrill 0.1

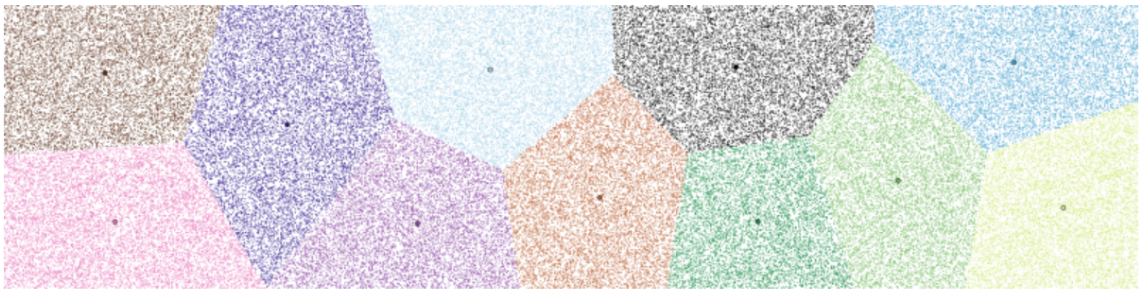
Main Page | Related Pages | Modules | Namespaces | Classes | Files | Examples | Search

- Thrill
  - Thrill Documentation Overview
    - Getting Started
    - K-Means Tutorial
      - Step 1: Generate Random Points
      - Step 2: Pick Random Centers and C
      - Step 3: ReduceByKey to Calculate N
      - Step 4: Iteration!
      - Step 5: Input and Output
      - Bonus Step 6: Boost Qi
    - Thrill Layer Architecture
    - Coding Style Guide
  - Modules
  - Namespaces
  - Classes
  - Files
  - Examples

Step 1: Generate Random Points

Welcome to the first step in the Thrill k-means tutorial. This tutorial will show how to implement the k-means clustering algorithm (Lloyd's algorithm) in Thrill.

The algorithm works as follows: Given a set of d-dimensional points, select k initial cluster center points at random. 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. This will be repeated a constant number of iterations.



We will implement this algorithm in Thrill, and only work with two-dimensional points for simplicity. Furthermore, we will hard-code many constants to make the code easier to understand.

In this step 1, let us start with generating random 2-dimensional points and outputting them for debugging.

We first need a Point class to represent the points. We may add some calculation functions to it later on.

```
//! A 2-dimensional point with double precision
struct Point {
    //! point coordinates
    double x, y;
};
```

For outputting the Point class, we need to add an `operator <<` for `std::ostream`, which is the standard way for

Thrill Documentation Overview | K-Means Tutorial | Generated on Tue Sep 20 2016 19:24:29 for Thrill by doxygen 1.8.5

Current and Future Work

- Open-Source at `http://project-thrill.org` and Github.
- High quality, very modern C++14 code.

Ideas for Future Work:

- Distributed rank()/select() and wavelet tree construction.
- Beyond `DIA<T>`? `Graph<V,E>`? `DenseMatrix<T>`?
- Fault tolerance? Go from p to $p-1$ workers?
- Communication efficient distributed operations for Thrill.
- Distributed functional programming language on top of Thrill.

Thank you for your attention!

Questions?