TASK-BASED PROGRAMMING MODELS FOR SCALABLE ALGORITHMS

Application to matrix multiplication

E. Agullo, A. Buttari, A. Guermouche, J. Herrmann, A. Jego

June 17, 2022

INTRODUCTION

Algorithms, architectures and programming models

Architectures are becoming increasingly **large**, **heterogeneous** and **complex**. This has two effects:

- Algorithms have to handle communications to achieve high scalability
  - Communication/computation overlap
  - Communication-avoiding algorithms
- Software should deliver a reliable abstraction to build mathematical software
  - parallel programming models
  - runtime systems

Are currently available parallel programming models and runtimes expressive enough to implement scalable algorithms in an efficient, portable and maintainable way?

Algorithms, architectures and programming models

Architectures are becoming increasingly **large**, **heterogeneous** and **complex**. This has two effects:

- Algorithms have to handle communications to achieve high scalability
  - Communication/computation overlap
  - Communication-avoiding algorithms

- Software should deliver a reliable abstraction to build mathematical software
  - parallel programming models
  - runtime systems

Are currently available parallel programming models and runtimes expressive enough to implement scalable algorithms in an efficient, portable and maintainable way?

Algorithms, architectures and programming models

Architectures are becoming increasingly **large**, **heterogeneous** and **complex**. This has two effects:

- Algorithms have to handle communications to achieve high scalability
  - Communication/computation overlap
  - Communication-avoiding algorithms
- Software should deliver a reliable abstraction to build mathematical software
  - parallel programming models
  - runtime systems

Are currently available parallel programming models and runtimes expressive enough to implement scalable algorithms in an efficient, portable and maintainable way?

Algorithms, architectures and programming models

Architectures are becoming increasingly **large**, **heterogeneous** and **complex**. This has two effects:

- Algorithms have to handle communications to achieve high scalability
  - Communication/computation overlap
  - Communication-avoiding algorithms
- Software should deliver a reliable abstraction to build mathematical software
  - parallel programming models
  - runtime systems

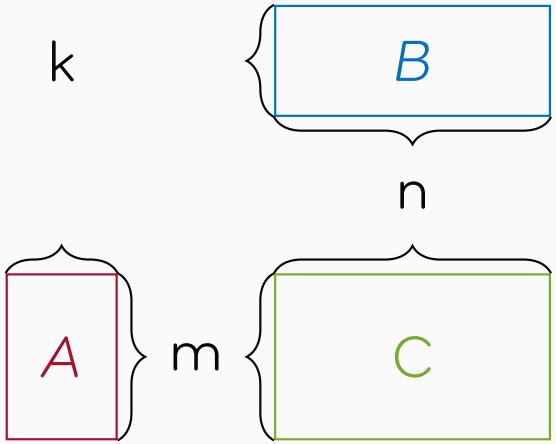
Are currently available parallel programming models and runtimes expressive enough to implement scalable algorithms in an efficient, portable and maintainable way?

Algorithms, architectures and programming models

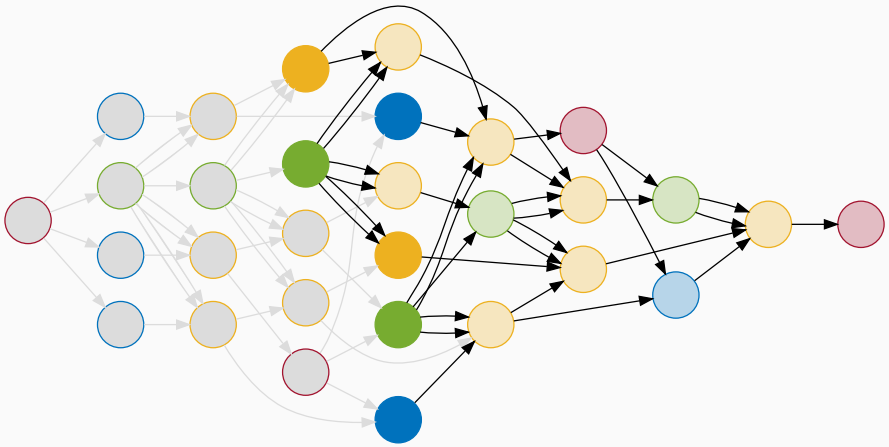
We take the following examples to answer this large question

- Algorithms: scalable dense matrix multiplication (GEMM) algorithms such as SUMMA and 2.5D SUMMA

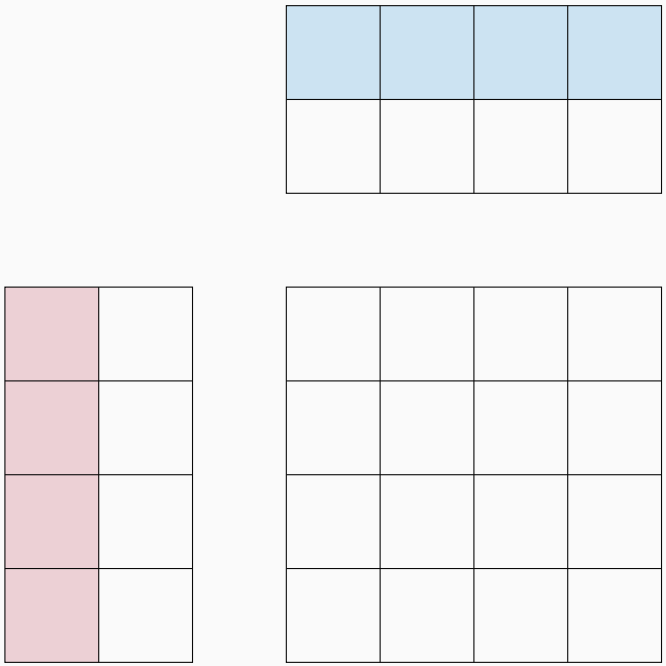
$$C = \alpha AB + \beta C$$



- Software: task-based parallel programming paradigm through the sequential task flow (STF) model



BACKGROUND

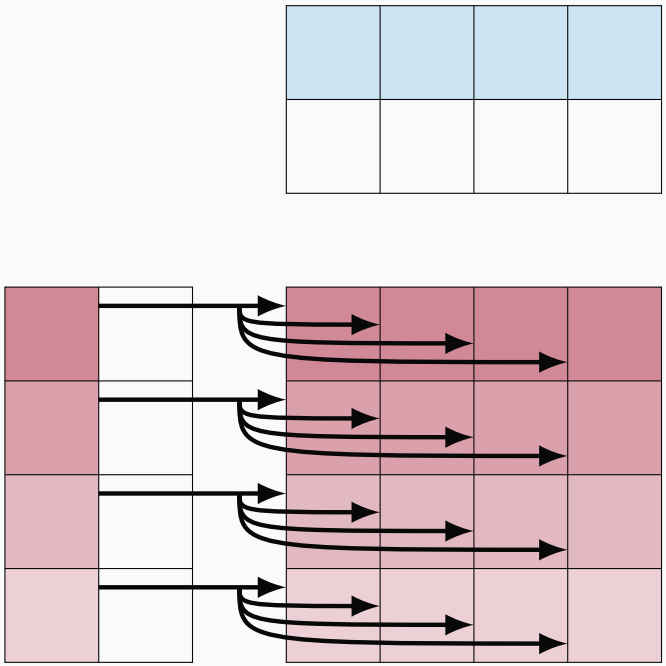


2D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Workload mapping**
  - owner of C computes
- **Communications**
  - row-wise broadcast of A
  - column-wise broadcast of B

```
! I am r × c
do l=1, k
   forall i, i%p==r :
      bcast(A_{i,l}, r×*)
   forall j, j%q==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall j, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C_{i,j})
end do
```

Geijn and Watts, 1997

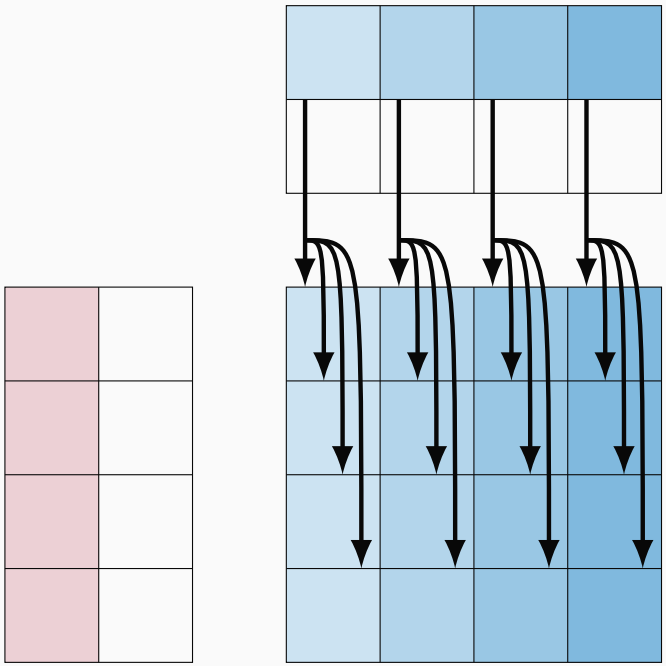


2D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Workload mapping**
  - owner of C computes
- **Communications**
  - row-wise **broadcast** of A
  - column-wise broadcast of B

```
! I am r × c
do l=1, k
   forall i, i%p==r :
      bcast(A_{i,l}, r×*)
   forall j, j%q==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall j, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C_{i,j})
end do
```

Geijn and Watts, 1997

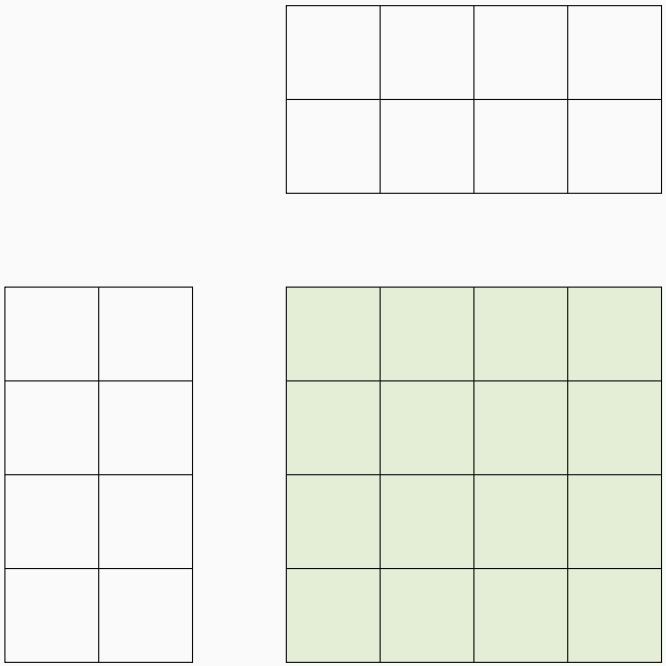


2D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Workload mapping**
  - owner of C computes
- **Communications**
  - row-wise broadcast of A
  - column-wise **broadcast** of B

```
! I am r × c
do l=1, k
   forall i, i%p==r :
      bcast(A_{i,l}, r×*)
   forall j, j%q==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall j, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C_{i,j})
end do
```

Geijn and Watts, 1997

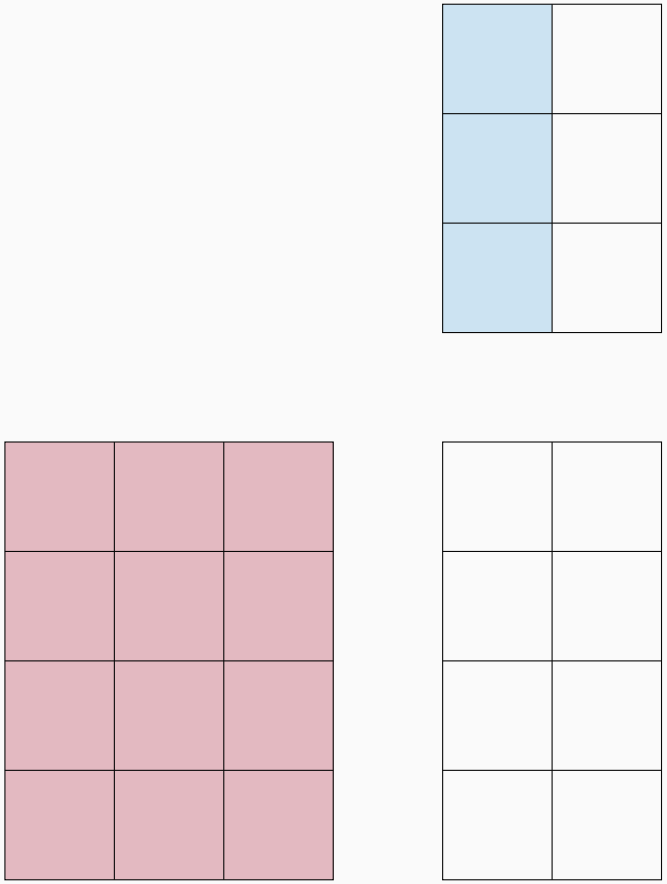


2D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Workload mapping**
  - owner of C **computes**
- **Communications**
  - row-wise broadcast of A
  - column-wise broadcast of B

```
! I am r × c
do l=1, k
   forall i, i%p==r :
      bcast(A_{i,l}, r×*)
   forall j, j%q==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall j, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C_{i,j})
end do
```

Geijn and Watts, 1997

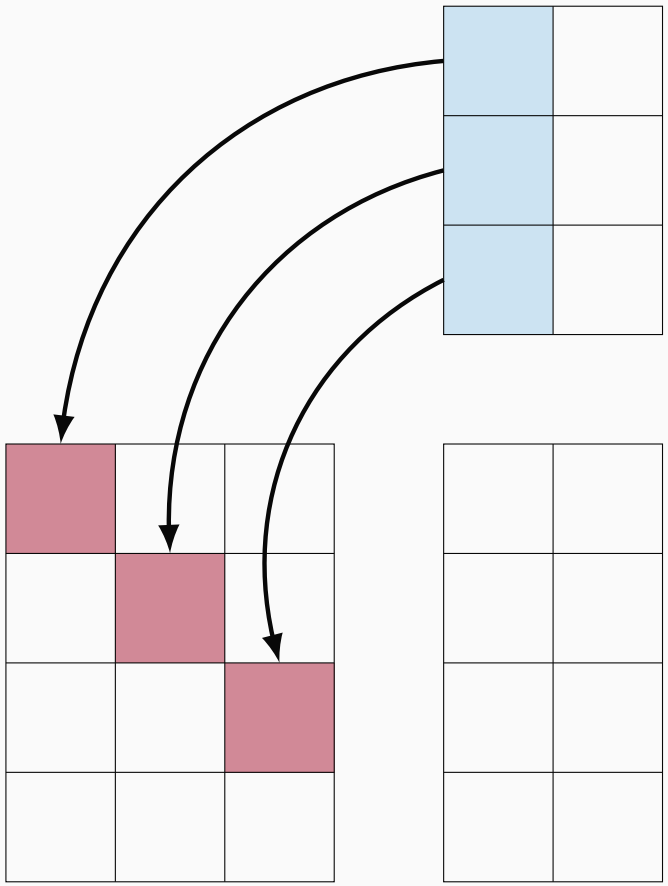


2D A-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Worload mapping**
  - owner of A computes
- **Communications**
  - column-wise scatter+broadcast of B
  - row-wise reduction of C

```
! I am r × c
do j=1, n
   forall l, l%p==r.and.l%q==c :
      recv(B_{l,j}, l%p×j%q)
   forall l, l%p==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall l, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C^r_{i,i})
      reduce(C^r_{i,i}, i%r×j%q)
end do
```

Geijn and Watts, 1997

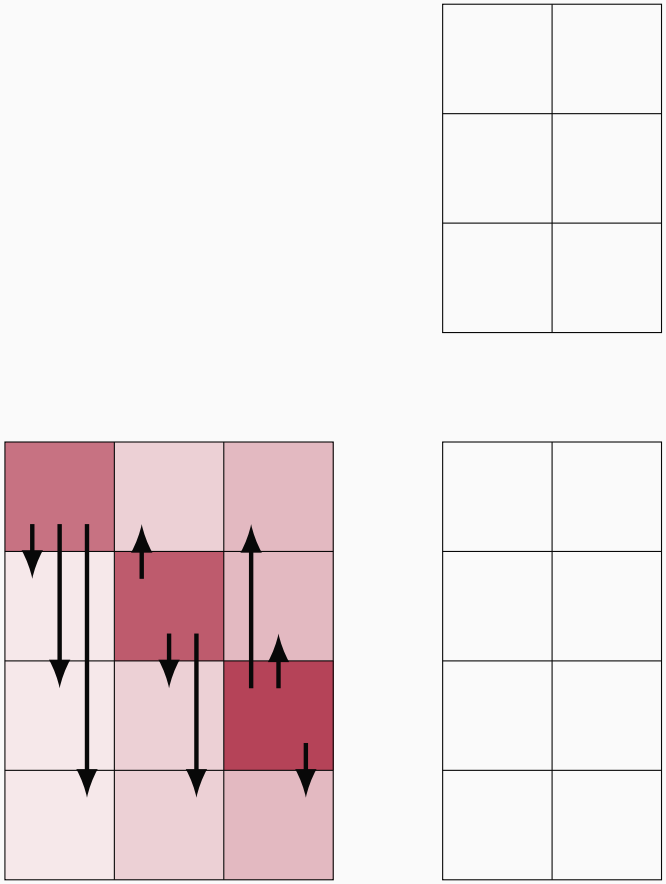


2D A-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Worload mapping**
  - owner of A computes
- **Communications**
  - column-wise **scatter**+broadcast of B
  - row-wise reduction of C

```
! I am r × c
do j=1, n
   forall l, l%p==r.and.l%q==c :
      recv(B_{l,j}, l%p×j%q)
   forall l, l%p==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall l, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C^r_{i,i})
      reduce(C^r_{i,i}, i%r×j%q)
end do
```

Geijn and Watts, 1997

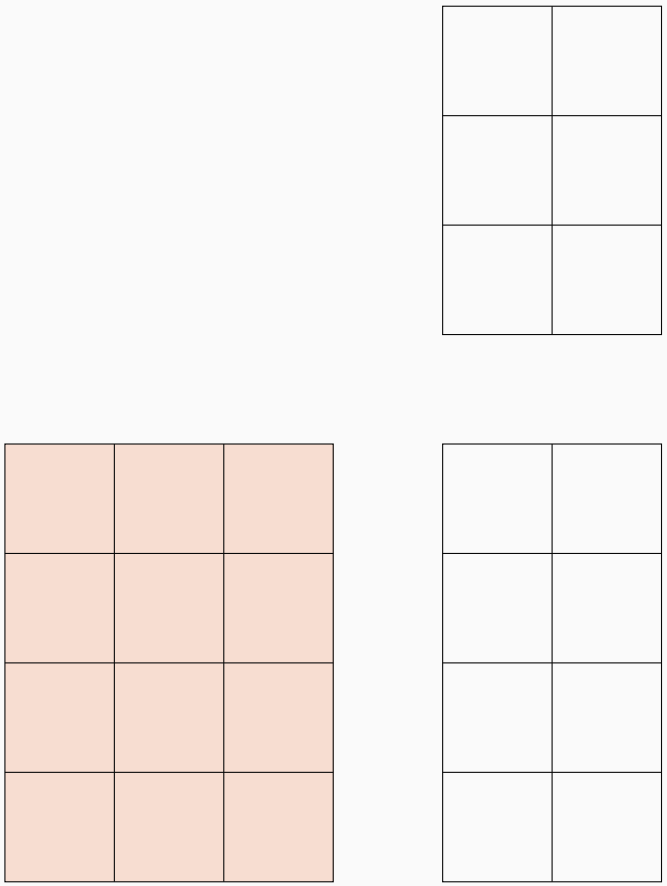


2D A-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Worload mapping**
  - owner of A computes
- **Communications**
  - column-wise scatter+**broadcast** of B
  - row-wise reduction of C

```
! I am r × c
do j=1, n
   forall l, l%p==r.and.l%q==c :
      recv(B_{l,j}, l%p×j%q)
   forall l, l%p==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall l, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C^r_{i,j})
      reduce(C^r_{i,j}, i%r×j%q)
end do
```

Geijn and Watts, 1997

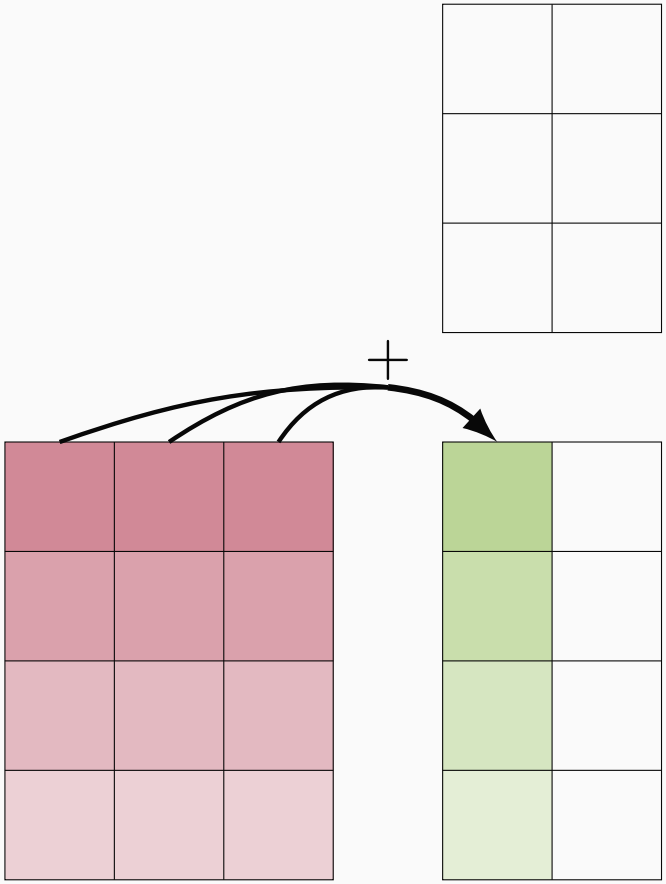


2D A-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Worload mapping**
  - owner of A **computes**
- **Communications**
  - column-wise scatter+broadcast of B
  - row-wise reduction of C

```
! I am r × c
do j=1, n
   forall l, l%p==r.and.l%q==c :
      recv(B_{l,j}, l%p×j%q)
   forall l, l%p==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall l, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C^r_{i,i})
      reduce(C^r_{i,i}, i%r×j%q)
end do
```

Geijn and Watts, 1997

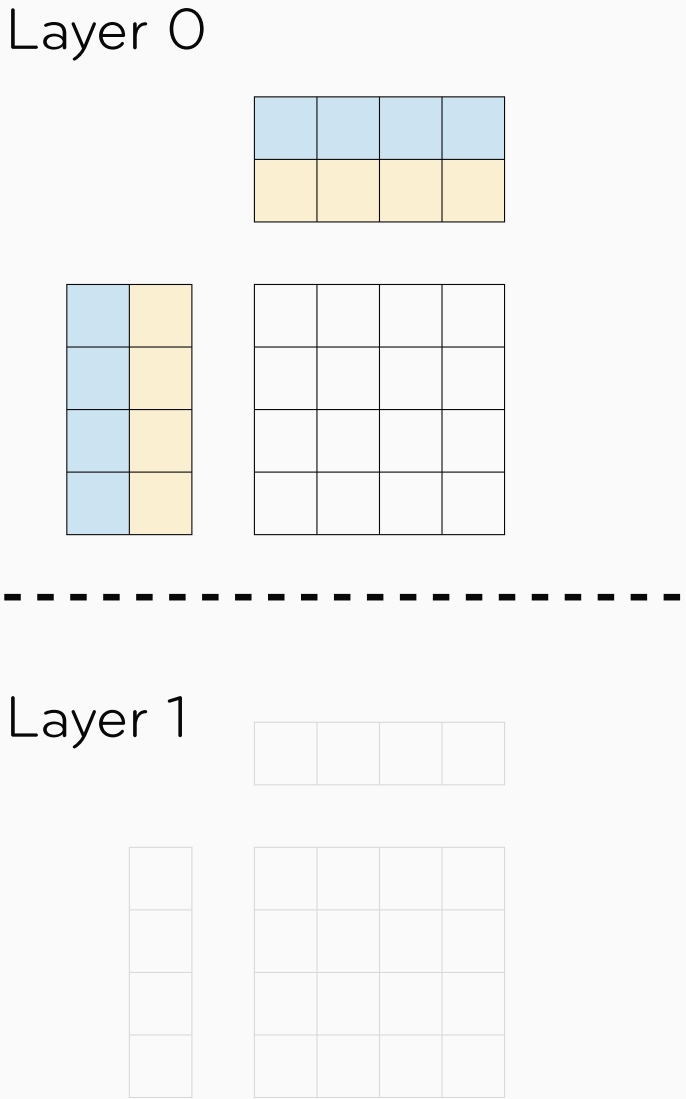


2D A-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic
- **Worload mapping**
  - owner of A computes
- **Communications**
  - column-wise scatter+broadcast of B
  - row-wise **reduction** of C

```
! I am r × c
do j=1, n
   forall l, l%p==r.and.l%q==c :
      recv(B_{l,j}, l%p×j%q)
   forall l, l%p==c :
      bcast(B_{l,j}, *×c)
   forall i, i%p==r :
      forall l, j%q==c :
         call gemm(A_{i,l}, B_{l,j}, C^r_{i,i})
      reduce(C^r_{i,i}, i%r×j%q)
end do
```

Geijn and Watts, 1997

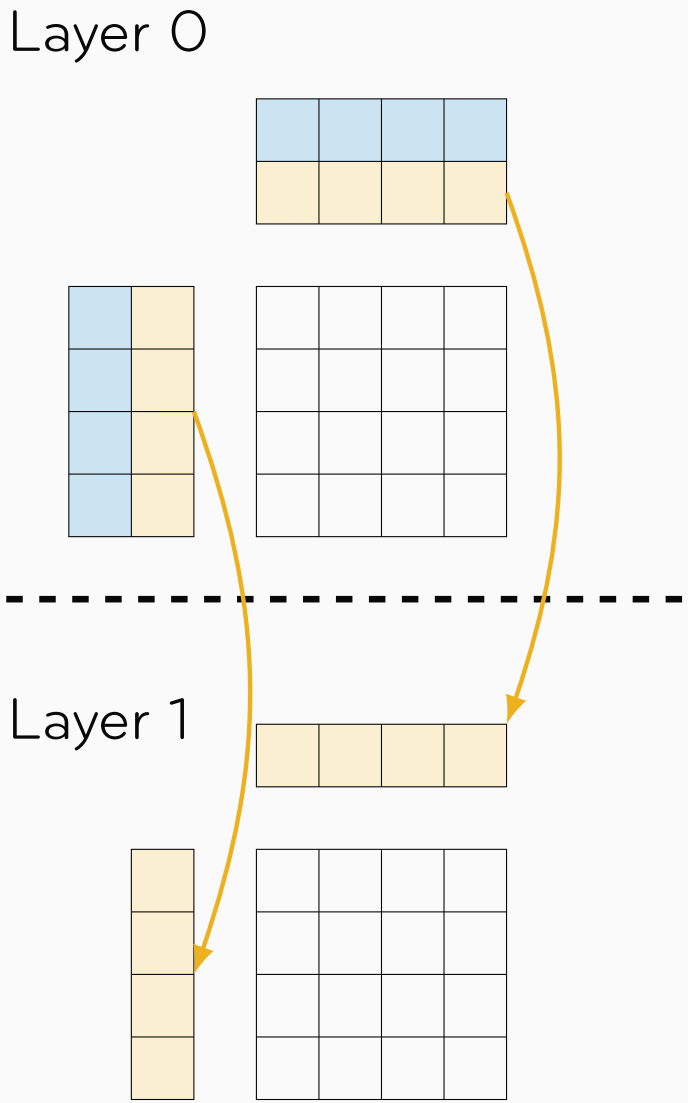


2.5D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic on 0^{th} layer
- **Workload mapping**
  - Not trivially related to data mapping
- **Communications**
  - row-wise scatter+broadcast of A
  - column-wise scatter+broadcast of B
  - aisle-wise reduction of C

```
! I am r × c × h
scatter A_{i,j} on layer j/h
scatter B_{i,j} on layer i/h

forall layer=1,h:
   call SUMMA_2DC(A[:, (layer : layer + 1) * k/h],
                  B[(layer : layer + 1) * k/h, :], C^h)

forall i, i%p==r :
  forall j, j%q==c :
    reduce(C^h_{i,j}, r×c×0)
```

Irony, Toledo, and Tiskin, 2004

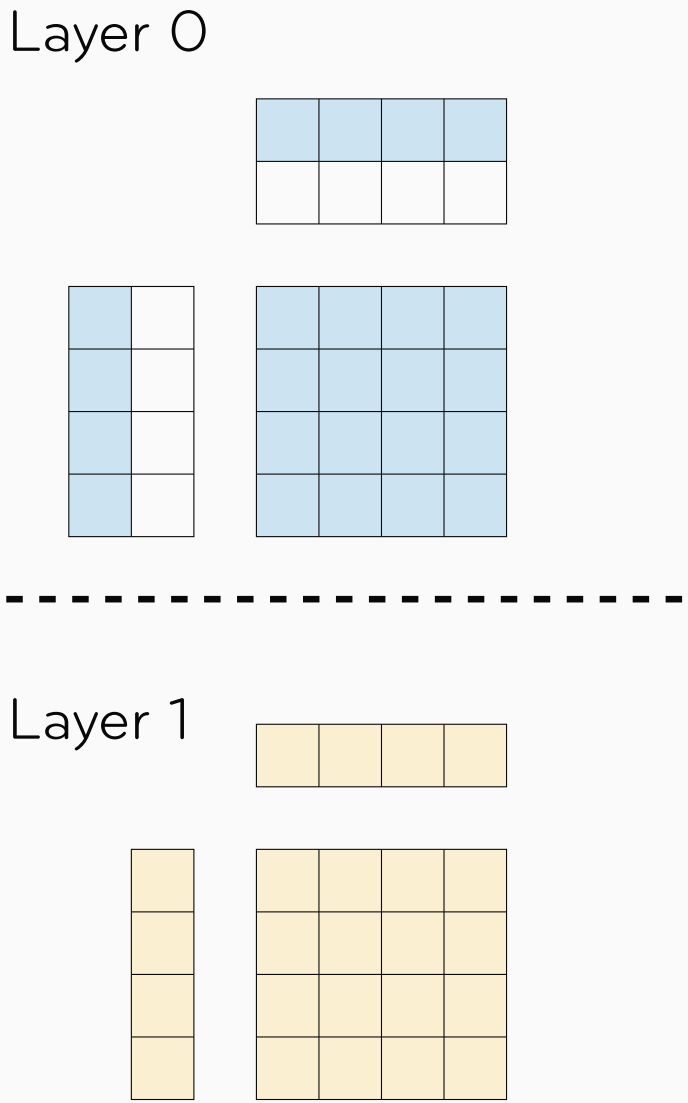


2.5D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic on 0^{th} layer
- **Workload mapping**
  - Not trivially related to data mapping
- **Communications**
  - row-wise **scatter**+broadcast of A
  - column-wise **scatter**+broadcast of B
  - aisle-wise reduction of C

```
! I am r × c × h
scatter A_{i,j} on layer j/h
scatter B_{i,j} on layer i/h

forall layer=1,h:
   call SUMMA_2DC(A[:, (layer : layer + 1) * k/h],
                  B[(layer : layer + 1) * k/h, :], C^h)

forall i, i%p==r :
  forall j, j%q==c :
    reduce(C^h_{i,j}, r×c×0)
```

Irony, Toledo, and Tiskin, 2004

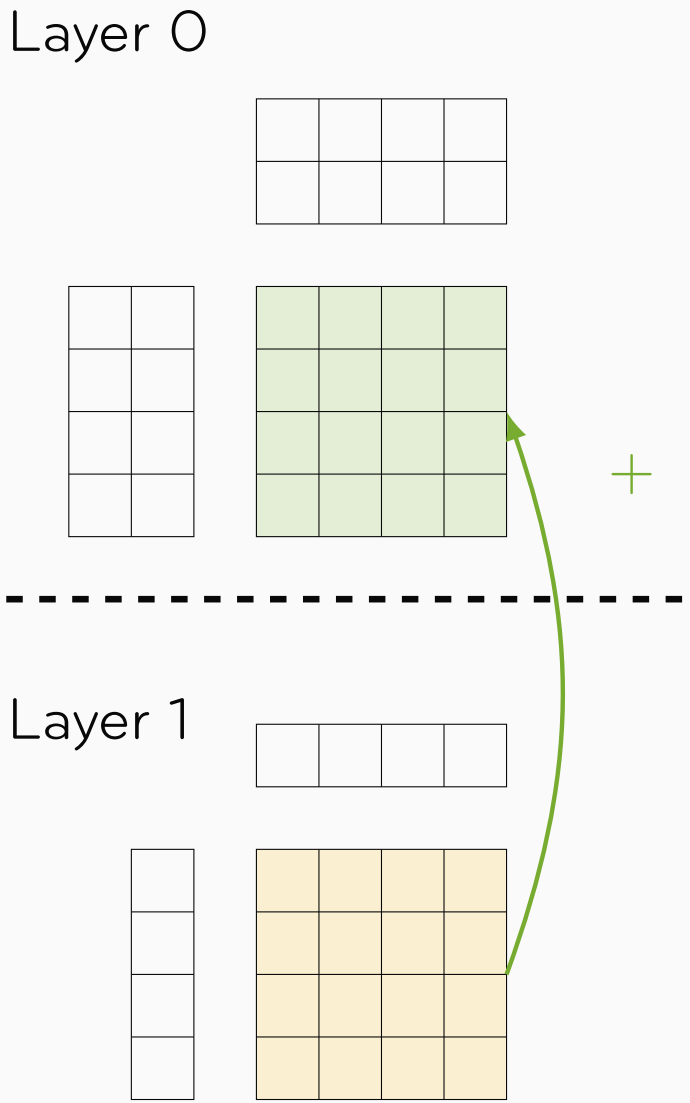


2.5D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic on 0^{th} layer
- **Workload mapping**
  - Not trivially related to data mapping
- **Communications**
  - row-wise scatter+**broadcast** of A
  - column-wise scatter+**broadcast** of B
  - aisle-wise reduction of C

```
! I am r × c × h
scatter A_{i,j} on layer j/h
scatter B_{i,j} on layer i/h

forall layer=1,h:
   call SUMMA_2DC(A[:, (layer : layer + 1) * k/h],
                  B[(layer : layer + 1) * k/h, :], C^h)

forall i, i%p==r :
  forall j, j%q==c :
    reduce(C^h_{i,j}, r×c×0)
```

Irony, Toledo, and Tiskin, 2004



2.5D C-stationary SUMMA

- **Data mapping**
  - 2D block-cyclic on 0^{th} layer
- **Workload mapping**
  - Not trivially related to data mapping
- **Communications**
  - row-wise scatter+broadcast of A
  - column-wise scatter+broadcast of B
  - aisle-wise **reduction** of C

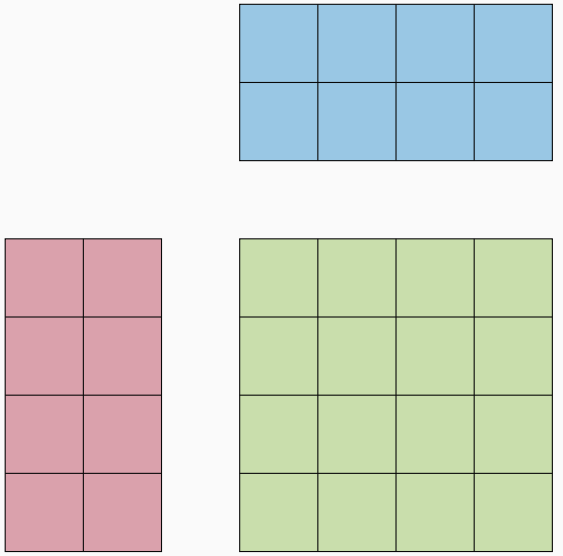
```
! I am r × c × h
scatter A_{i,j} on layer j/h
scatter B_{i,j} on layer i/h

forall layer=1,h:
   call SUMMA_2DC(A[:, (layer : layer + 1) * k/h],
                  B[(layer : layer + 1) * k/h, :], C^h)

forall i, i%p==r :
  forall j, j%q==c :
    reduce(C^h_{i,j}, r×c×0)
```

Irony, Toledo, and Tiskin, 2004

The sequential task flow (STF) Model

Sequential GEMM

```
do i=1,m
  do j=1,n
    do l=1,k
      call gemm(
                    A(i,l),
                    B(l,j),
                    C(i,j) )
    end do
  end do
end do
```



Shared-memory STF GEMM

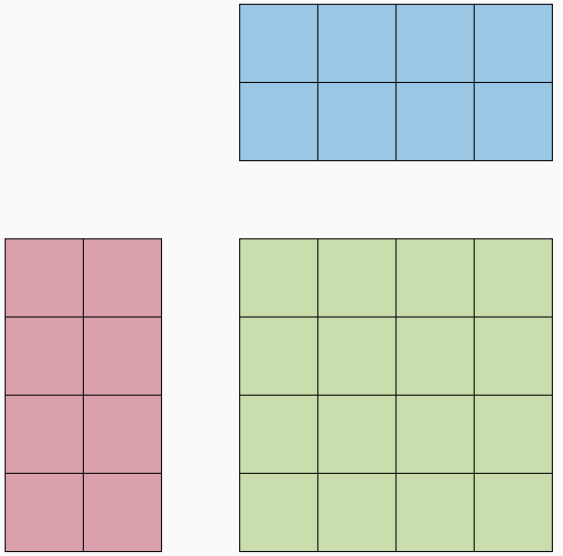
```
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task( gemm,
                    A(i,l):R,
                    B(l,j):R,
                    C(i,j):RW)
    end do
  end do
end do
```


DAG of GEMM

The sequential task flow (STF) Model

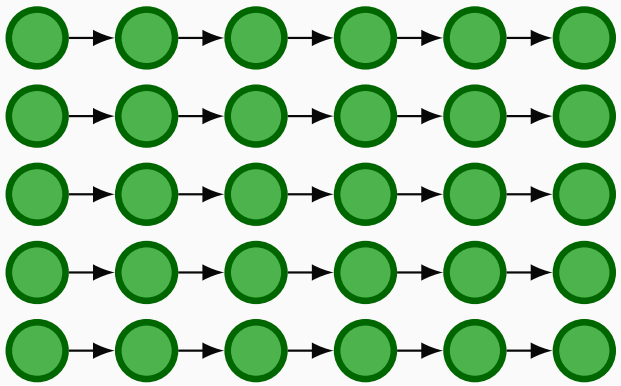
Sequential GEMM

```
do i=1,m
  do j=1,n
    do l=1,k
      call gemm(
                  A(i,l),
                  B(l,j),
                  C(i,j) )
    end do
  end do
end do
```



Shared-memory STF GEMM

```
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task( gemm,
                  A(i,l):R,
                  B(l,j):R,
                  C(i,j):RW)
    end do
  end do
end do
```



DAG of GEMM

Objectives

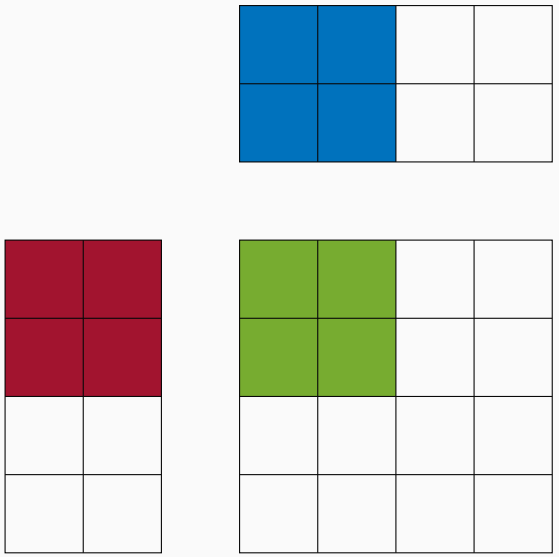
- Can we use the **STF model** to express a **scalable distributed-memory GEMM** algorithm ?

- How **efficient** is the resulting expression compared to reference libraries ?

DISTRIBUTED-MEMORY STF GEMM ALGORITHM

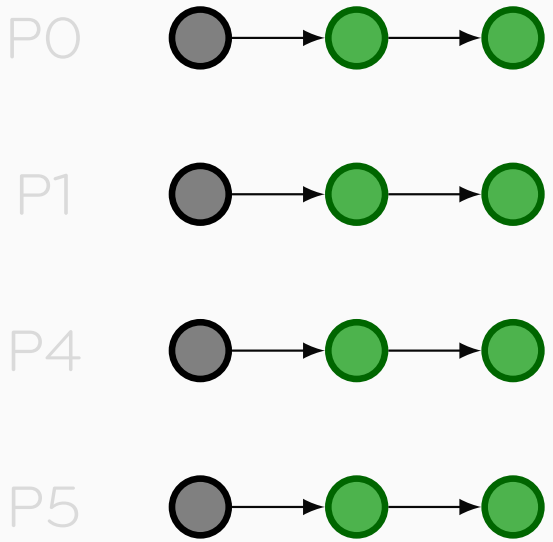
Baseline STF model

```
!
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW)
    end do
  end do
end do
```



Baseline model (M0)[a]

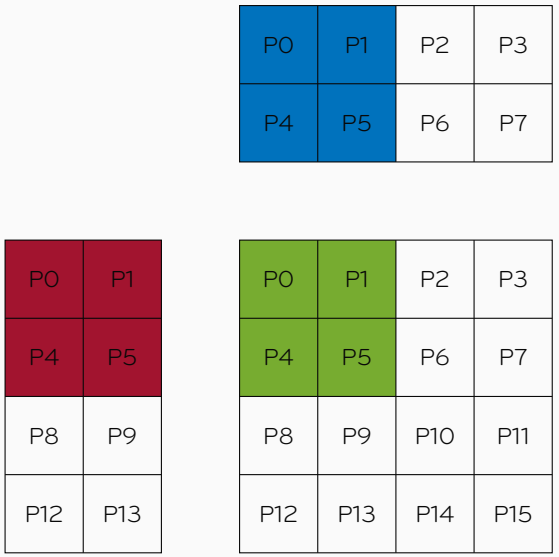
- M0-0 Data mapping
- M0-1 Task mapping **inferred** from DM (RW)
- M0-2 Point-to-point comms. **inferred** from TM



[a] Agullo, Aumage, Faverge, Furmento, Pruvost, Sergent, and Thibault, 2017.

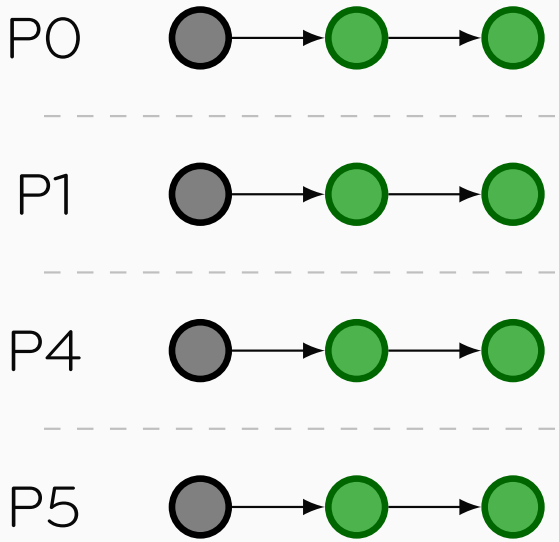
Baseline STF model

```
! data_map: A,B,C distributed on a grid
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW)
    end do
  end do
end do
```



Baseline model (M0)[a]

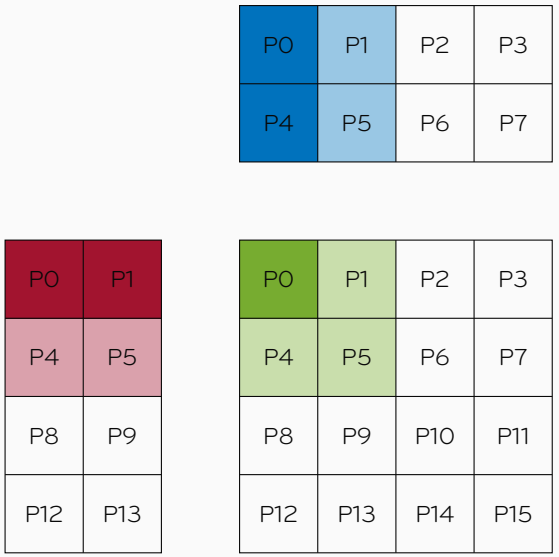
- M0-0 Data mapping
- M0-1 Task mapping **inferred** from DM (RW)
- M0-2 Point-to-point comms. **inferred** from TM



[a] Agullo, Aumage, Faverge, Furmento, Pruvost, Sergent, and Thibault, 2017.

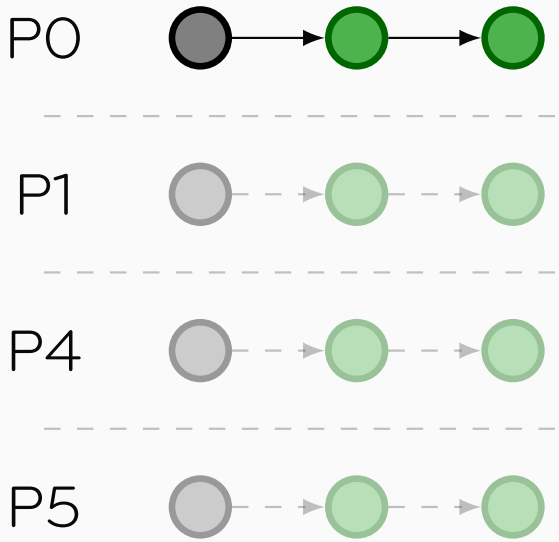
Baseline STF model

```
! data_map: A,B,C distributed on a grid
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW)
    end do
  end do
end do
```



Baseline model (M0)[a]

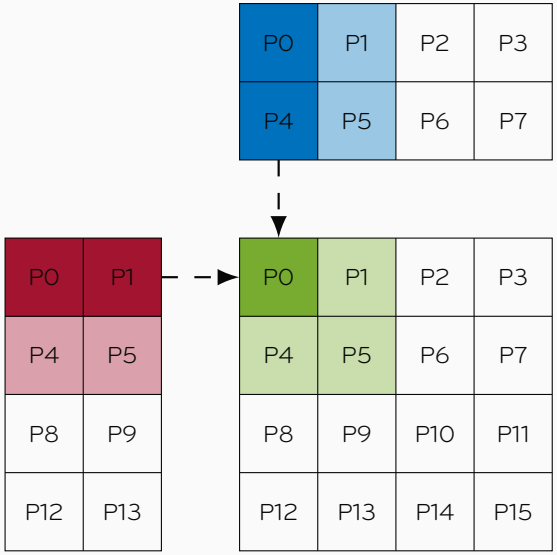
- M0-0 Data mapping
- M0-1 Task mapping **inferred** from DM (RW)
- M0-2 Point-to-point comms. **inferred** from TM



[a] Agullo, Aumage, Faverge, Furmento, Pruvost, Sergent, and Thibault, 2017.

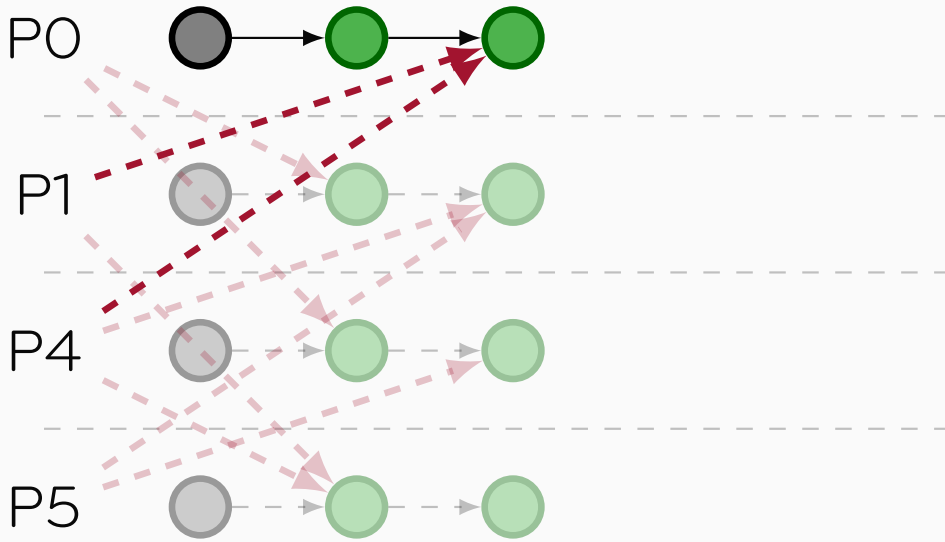
Baseline STF model

```
! data_map: A,B,C distributed on a grid
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW)
    end do
  end do
end do
```



Baseline model (M0)[a]

- M0-0 Data mapping
- M0-1 Task mapping **inferred** from DM (RW)
- M0-2 Point-to-point comms. **inferred** from TM



[a] Agullo, Aumage, Faverge, Furmento, Pruvost, Sergent, and Thibault, 2017.

Limits to the baseline model

The baseline STF model `M0` does not allow the expression of SUMMA algorithms

- Task mapping is bound to data mapping (`M0-1`)
- Reduction patterns are hard to make
- Point-to-point communications patterns are inefficient (`M0-2`)

We have identified key functionalities required in a STF model `MX`, extended from `M0`, to allow the expression of SUMMA algorithms.

Limits to the baseline model

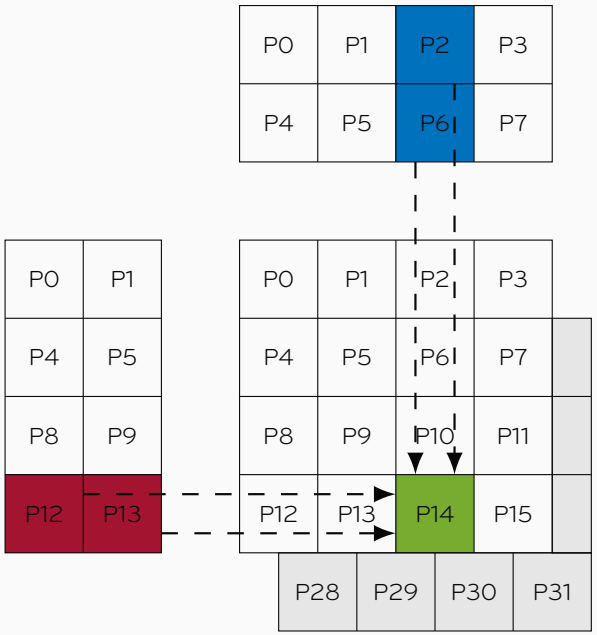
The baseline STF model `M0` does not allow the expression of SUMMA algorithms

- Task mapping is bound to data mapping (`M0-1`)
- Reduction patterns are hard to make
- Point-to-point communications patterns are inefficient (`M0-2`)

We have identified key functionalities required in a STF model `MX`, extended from `M0`, to allow the expression of SUMMA algorithms.

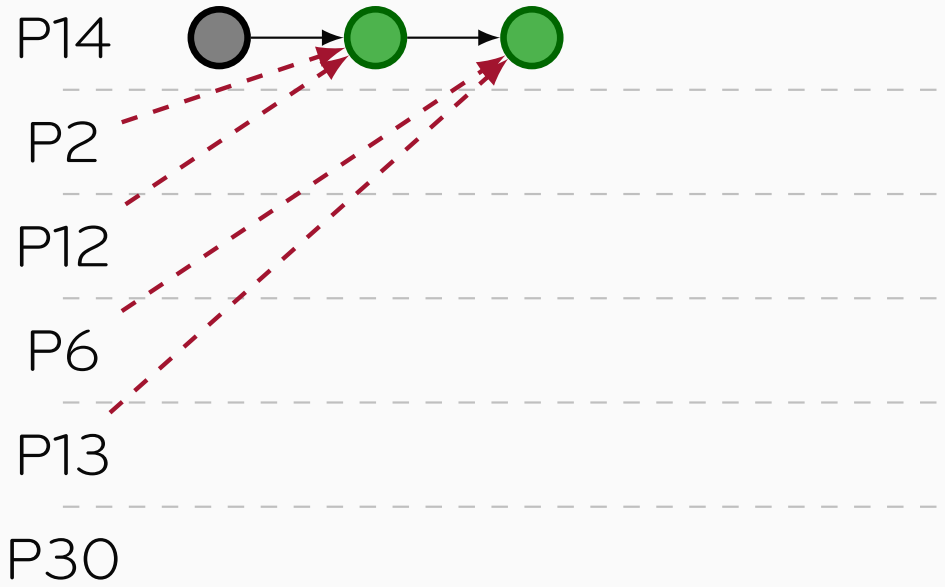
Key 1 – Task mapping

```
!data_map: A,B,C distributed on one layer
!
!

do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW)
    end do
  end do
end do
```



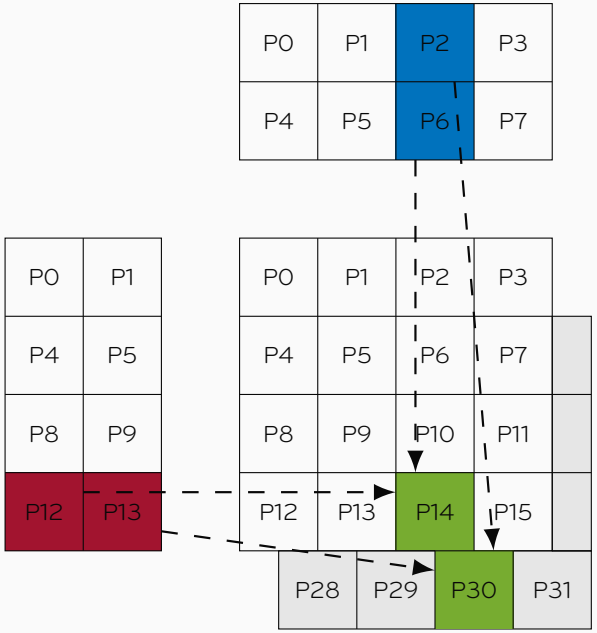
Advanced model MX

- M0-0 Data mapping
- M0-1 Implicit task mapping
- M0-2 Comms. inferred



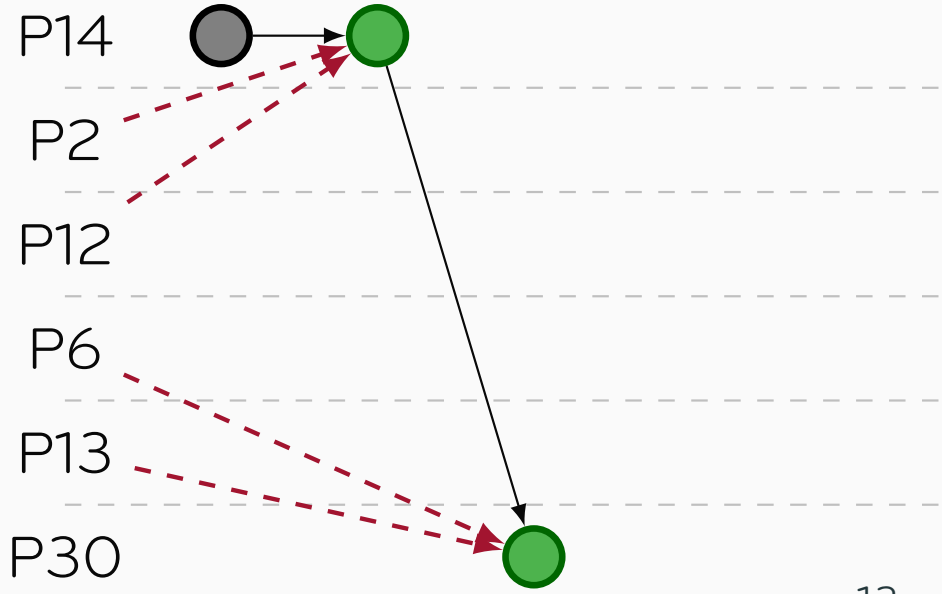
Key 1 – Task mapping

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!

do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW,
                       task_map[i,j,l])
    end do
  end do
end do
```



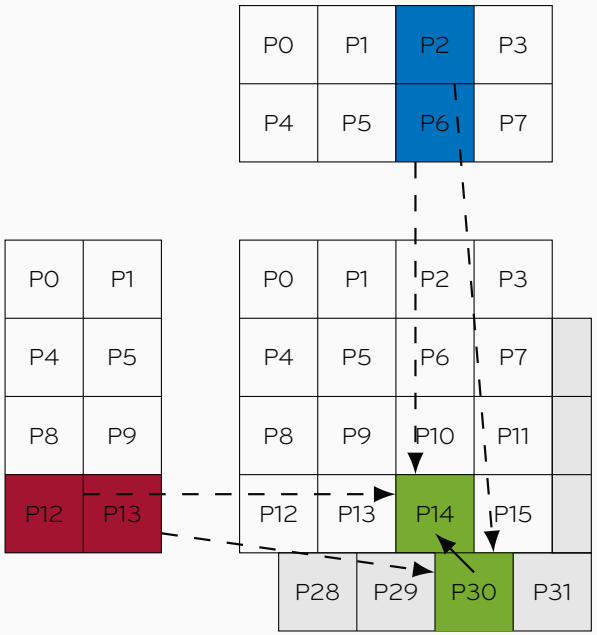
Advanced model MX

- M0-0 Data mapping
- MX-1 **Explicit** task mapping
- M0-2 Comms. inferred



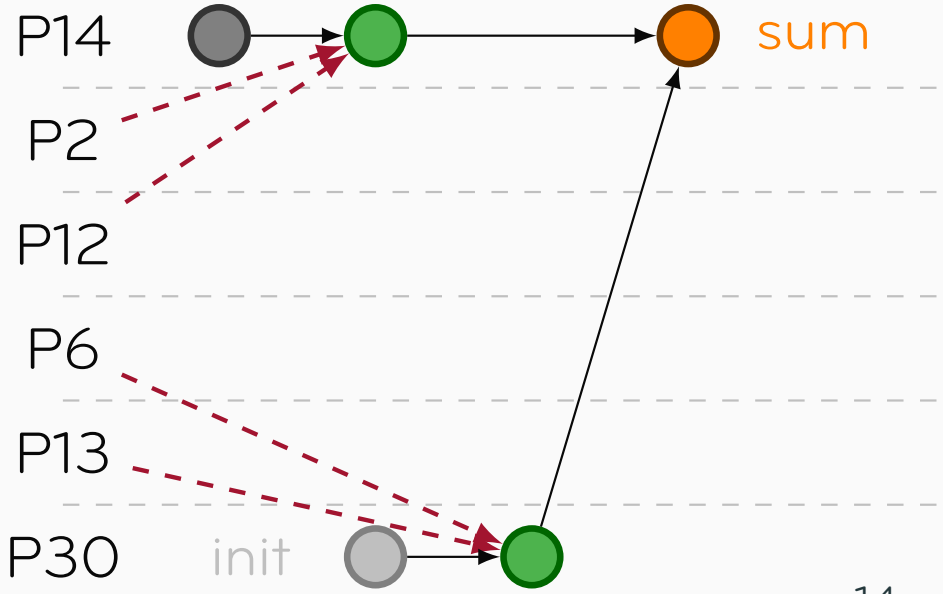
Key 2 – Reduction pattern

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!
call instantiate_local(C)
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RW,
                       task_map[i,j,l])
    end do
  end do
end do
call reduce_to_initial_layer(C)
```



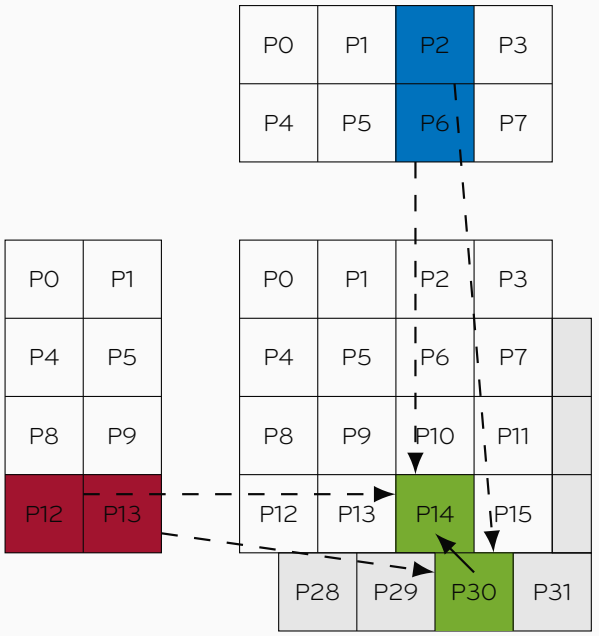
Advanced model MX

- MO-0 Data mapping
- MX-1 Explicit task mapping
- MO-2 Comms. inferred
- MO-3 Explicit reduction pattern



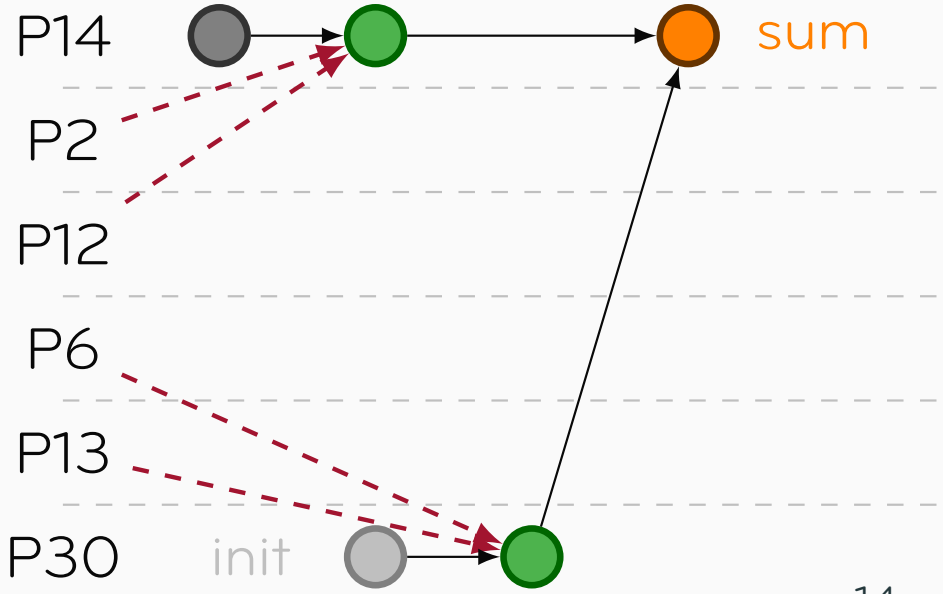
Key 2 – Reduction pattern

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!
call instantiate_local(C)
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RANK_REDUX,
                       task_map[i,j,l])
    end do
  end do
end do
call reduce_to_initial_layer(C)
```



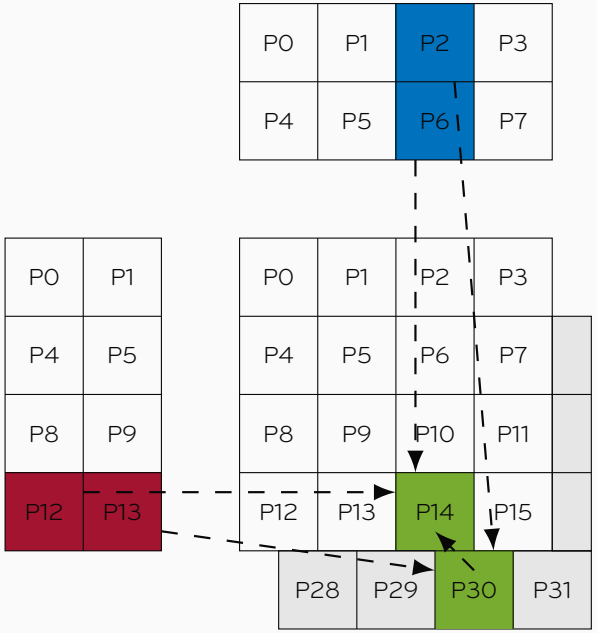
Advanced model MX

- MO-0 Data mapping
- MX-1 Explicit task mapping
- MO-2 Comms. inferred
- MO-3 Explicit reduction pattern



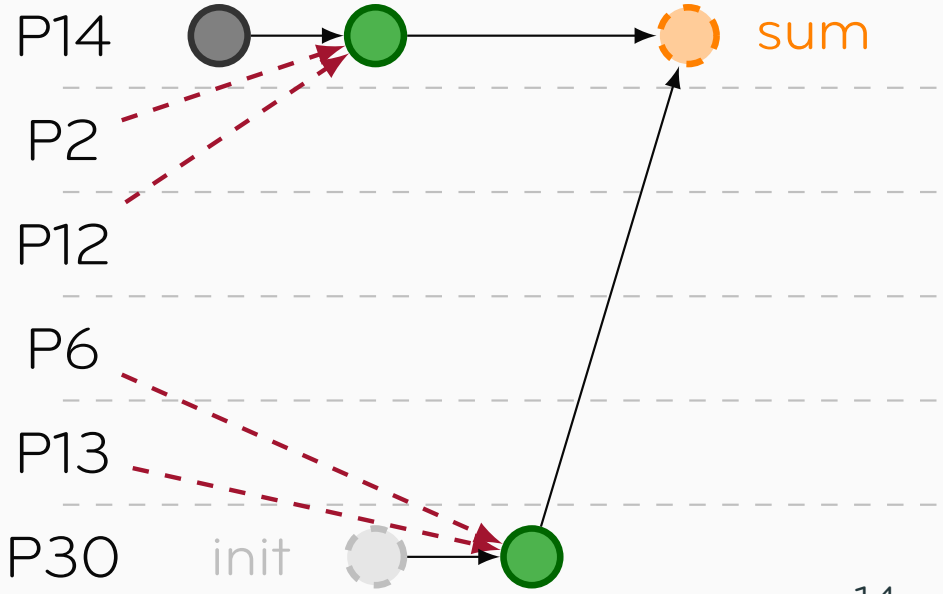
Key 2 – Reduction pattern

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!
call bind_methods(C, init, sum)
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RANK_REDUX,
                       task_map[i,j,l])
    end do
  end do
end do
```



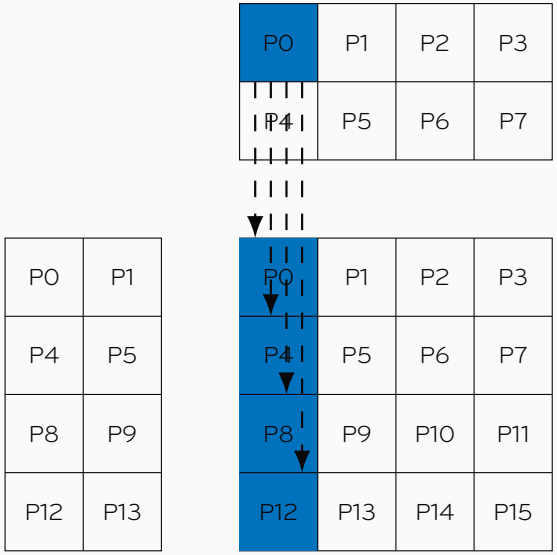
Advanced model MX

- MO-0 Data mapping
- MX-1 Explicit task mapping
- MO-2 Comms. inferred
- MX-3 **Implicit** reduction pattern



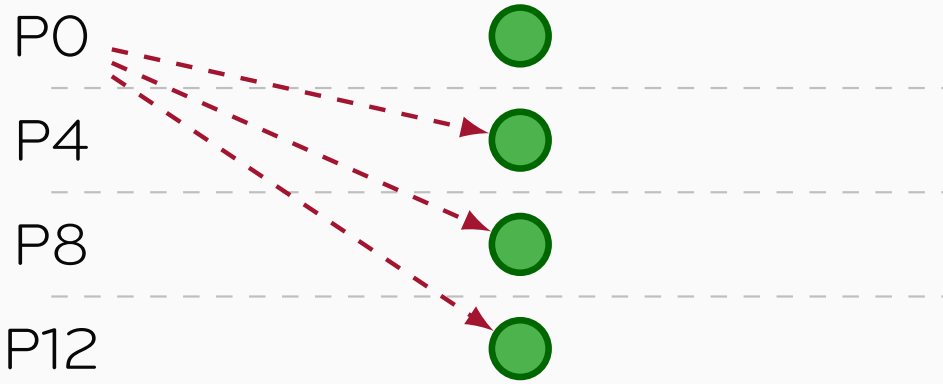
Key 3 – Communications

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!
call bind_methods(C, init, sum)
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RANK_REDUX,
                       task_map[i,j,l])
    end do
  end do
end do
```



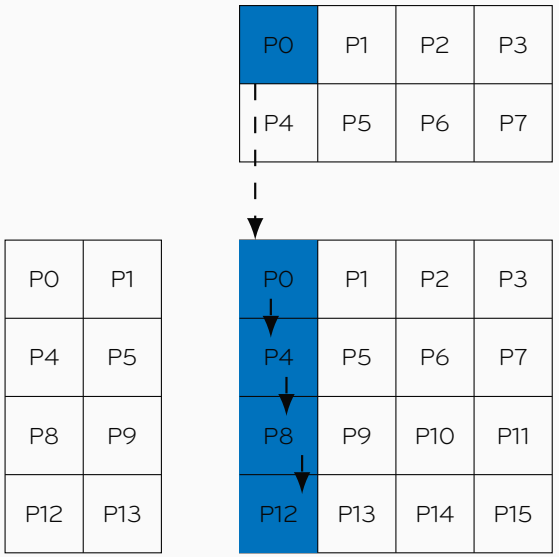
Advanced model MX

- MO-0 Data mapping
- MX-1 Explicit task mapping
- MO-2 Comms. inferred
- MX-3 Implicit reduction pattern



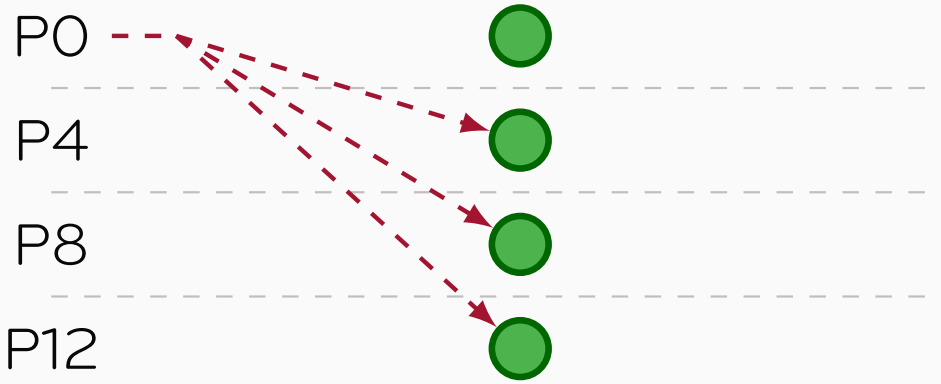
Key 3 – Communications

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!dyn_tree: type of broadcast tree
call bind_methods(C, init, sum)
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RANK_REDUX,
                       task_map[i,j,l])
    end do
  end do
end do
```



Advanced model MX

- MO-0 Data mapping
- MX-1 Explicit task mapping
- MX-2 **Coll. comms.** inferred [1]
- MX-3 Implicit reduction pattern



[1] Denis, Jeannot, Swartvagher, and Thibault, 2020.

Ideal expression

```
!data_map: A,B,C distributed on one layer
!task_map: executing node of A_{i,l}B_{l,j}
!dyn_tree: type of broadcast tree
call bind_methods(C, init, sum)
do i=1,m
  do j=1,n
    do l=1,k
      call insert_task(gemm,
                       A_{i,l}:R,
                       B_{l,j}:R,
                       C_{i,j}:RANK_REDUX,
                       task_map[i,j,l])
    end do
  end do
end do
```

- `task_map` can be set to cover any SUMMA variant
- Agnostic of:
  - the data distribution
  - the scheduling
  - the processing units

Advanced model `MX`

- `M0-0` Data mapping
- `MX-1` **Explicit** task mapping
- `MX-2` **Collective** communications **inferred** from task mapping
- `MX-3` **Implicit** reduction pattern

EXPERIMENTAL RESULTS

Experimental setup

- **Software**: GNU 9.3.0, MKL 21.3.0
  - `qr_mumps` **ours** – newMadeleine, 1 MPI / node
    - StarPU with advanced model `MX` + seq. BLAS
    - 2.5D A,B,C-stationary version
  - `scalapack` – OpenMPI 4.0.5, 2 cores / MPI
    - Synchronous MPI + seq. BLAS
    - 2D A,B,C-stationary
  - `slate` – OpenMPI 4.0.5, 1 MPI / node
    - MPI + OpenMP with `task` + seq. BLAS
    - 2D A,C-stationary with lookahead
  - `chameleon` – newMadeleine, 1 MPI / node
    - StarPU with explicit comms. management + seq. BLAS
    - 2D C-stationary pipelined with lookahead
- **Hardware**: Très Grand Centre de Calcul (TGCC)
  - **Irène - Skylake** Infiniband EDR, 4x links
    - Dual-processor Intel Skylake 8168 @ 2.7 GHz
    - 24 cores per processor, 192 GB DDR4 memory
  - **Irène - Rome** Infiniband HDR100, 2x links
    - Dual-processor AMD Rome (Epyc) @ 2.6 GHz
    - 64 cores per processor, 256 GB DDR4 memory

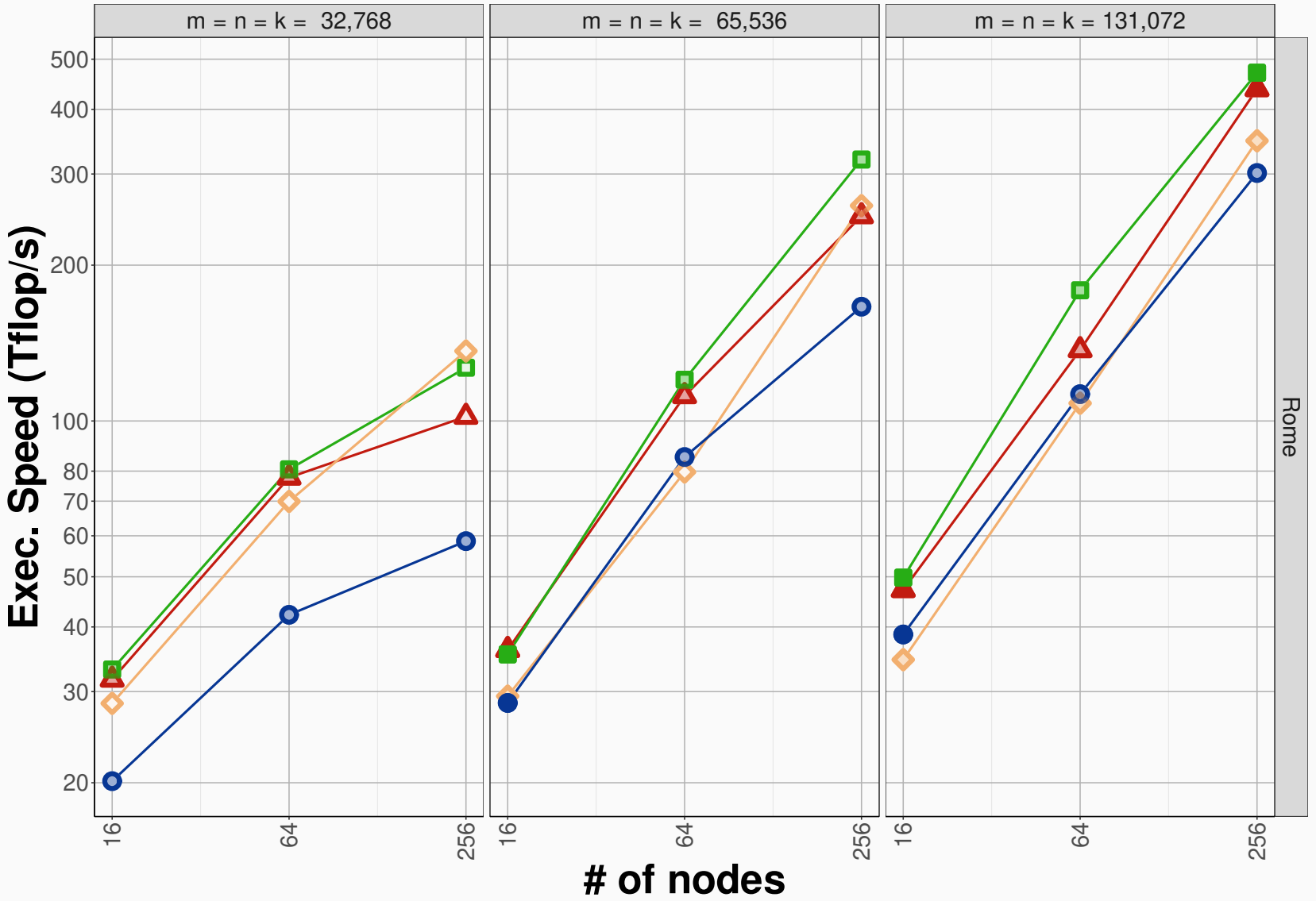
Experimental setup

- **Software**: GNU 9.3.0, MKL 21.3.0
  - `qr_mumps` **ours** – newMadeleine, 1 MPI / node
    - StarPU with advanced model `MX` + seq. BLAS
    - 2.5D A,B,C-stationary version
  - `scalapack` – OpenMPI 4.0.5, 2 cores / MPI
    - Synchronous MPI + seq. BLAS
    - 2D A,B,C-stationary
  - `slate` – OpenMPI 4.0.5, 1 MPI / node
    - MPI + OpenMP with `task` + seq. BLAS
    - 2D A,C-stationary with lookahead
  - `chameleon` – newMadeleine, 1 MPI / node
    - StarPU with explicit comms. management + seq. BLAS
    - 2D C-stationary pipelined with lookahead
- **Hardware**: Très Grand Centre de Calcul (TGCC)
  - **Irène - Skylake** Infiniband EDR, 4x links
    - Dual-processor Intel Skylake 8168 @ 2.7 GHz
    - 24 cores per processor, 192 GB DDR4 memory
  - **Irène - Rome** Infiniband HDR100, 2x links
    - Dual-processor AMD Rome (Epyc) @ 2.6 GHz
    - 64 cores per processor, 256 GB DDR4 memory

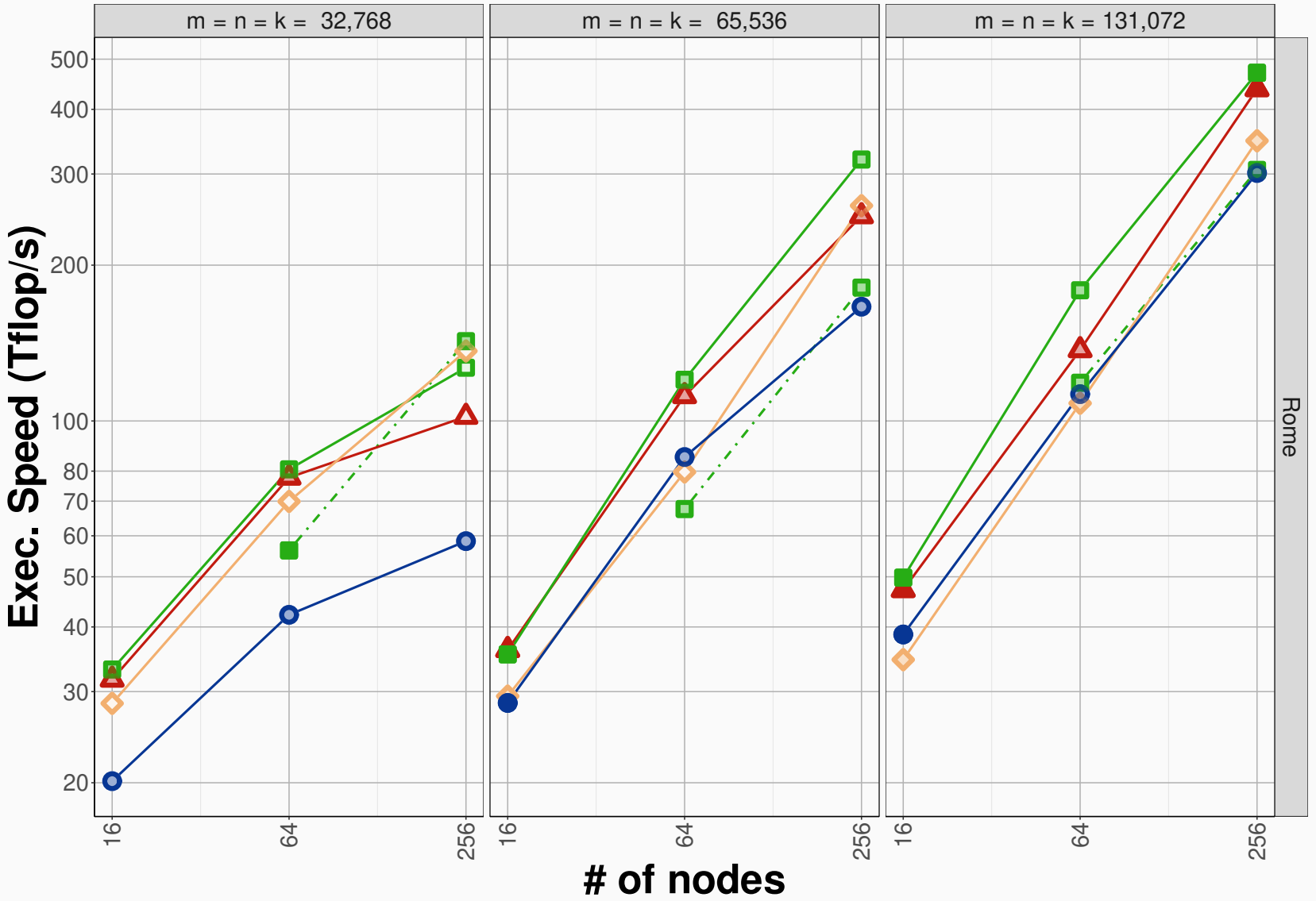
Experimental setup

- **Software**: GNU 9.3.0, MKL 21.3.0
  - `qr_mumps` **ours** – newMadeleine, 1 MPI / node
    - StarPU with advanced model `MX` + seq. BLAS
    - 2.5D A,B,C-stationary version
  - `scalapack` – OpenMPI 4.0.5, 2 cores / MPI
    - Synchronous MPI + seq. BLAS
    - 2D A,B,C-stationary
  - `slate` – OpenMPI 4.0.5, 1 MPI / node
    - MPI + OpenMP with `task` + seq. BLAS
    - 2D A,C-stationary with lookahead
  - `chameleon` – newMadeleine, 1 MPI / node
    - StarPU with explicit comms. management + seq. BLAS
    - 2D C-stationary pipelined with lookahead
- **Hardware**: Très Grand Centre de Calcul (TGCC)
  - **Irène - Skylake** Infiniband EDR, 4x links
    - Dual-processor Intel Skylake 8168 @ 2.7 GHz
    - 24 cores per processor, 192 GB DDR4 memory
  - **Irène - Rome** Infiniband HDR100, 2x links
    - Dual-processor AMD Rome (Epyc) @ 2.6 GHz
    - 64 cores per processor, 256 GB DDR4 memory

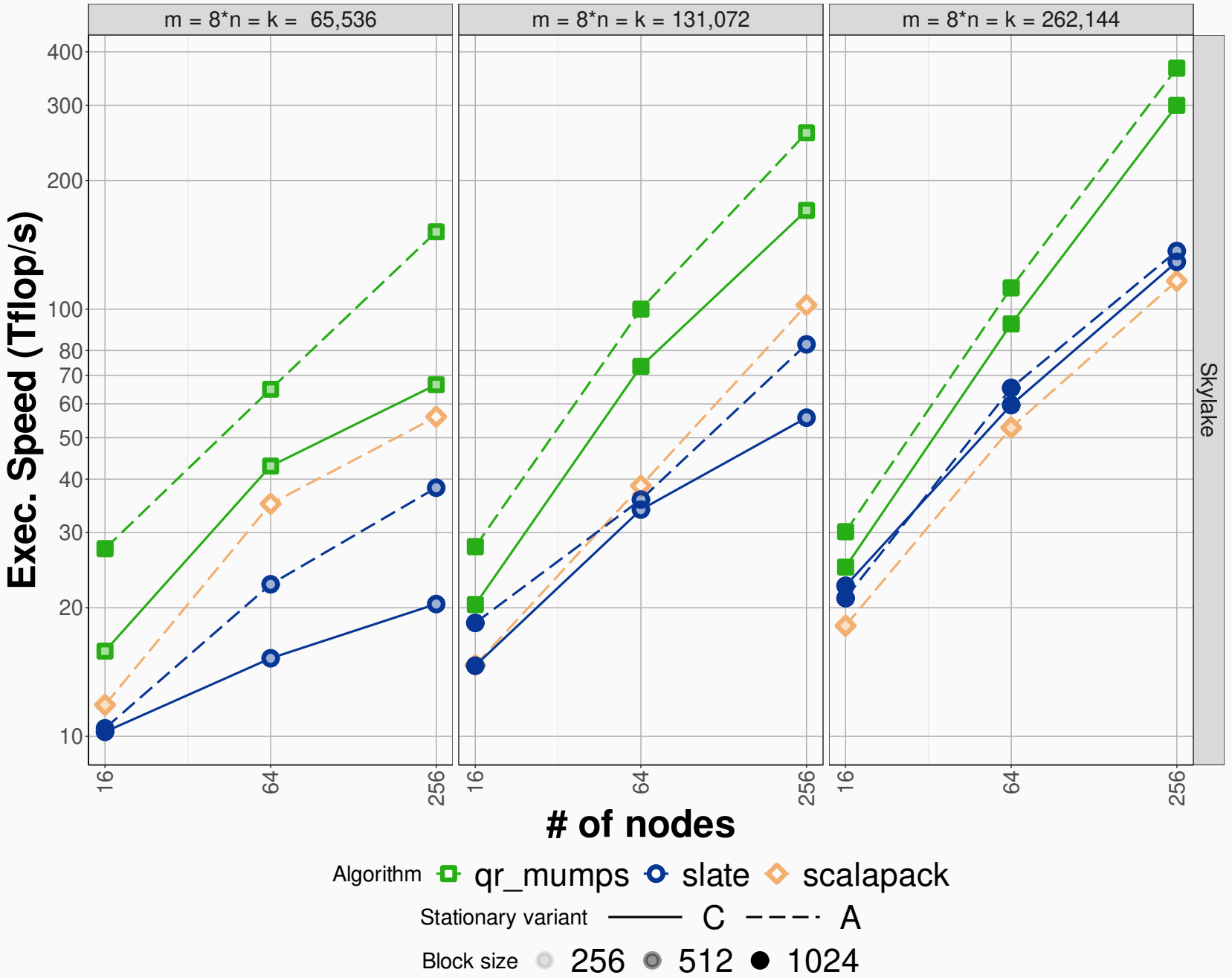
Rome (up to 32,768 cores) – square matrices, DGEMM



Rome (up to 32,768 cores) – square matrices, DGEMM



Skylake (up to 12,288 cores) – large A, DGEMM



CONCLUSIONS

Contributions

- Show **the STF programming model** can be suitably **extended** to **express** a complex distributed-memory algorithm
  - **Portable**, easily **maintainable**
  - Independant of matrices' distribution, workflow's scheduling
  - **Competitive** with reference libraries

Technical report: `https://hal.inria.fr/hal-03588491/`

On-going work

- Studying SYMM in the context of a scientific application
  - In collaboration with O. Beaumont, O. Coulaud, L. Eyraus-Dubois, A. Franc, R. Peressoni, F. Pruvost, M. Vérité
- Studying the scheduling of communication patterns in GEMM
  - In collaboration with O. Beaumont, A. Denis, L. Eyraus-Dubois, E. Jeannot, M. Vérité, P. Swartvagher

Future work

We want to address more distributed-memory scalable algorithms

- Dense **2.5D** Cholesky, LU **factorizations**
- Sparse factorizations

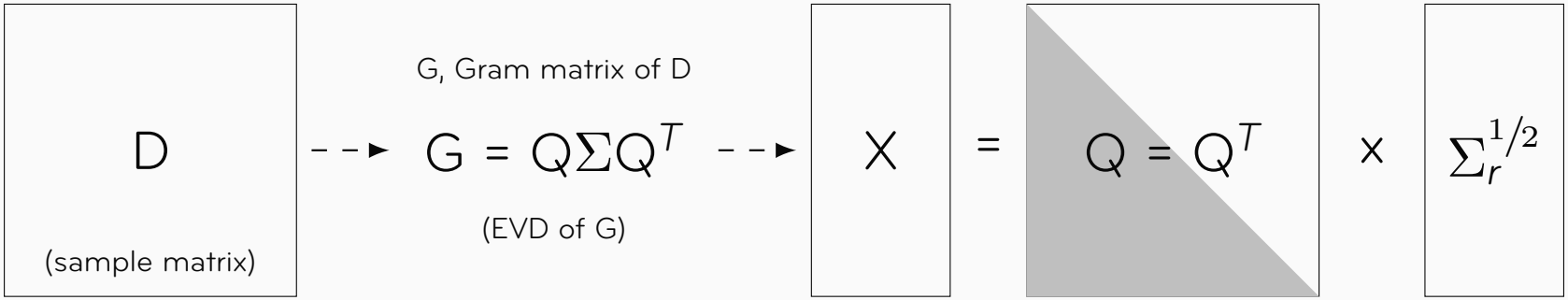
However, we need further **extensions** to the current programming model

- **Replication of computation** as required by **state-of-the-art 2.5D** algorithms
- **All-reduce pattern** to ensure **stability** through tournament pivoting
- **Memory-awareness** to ensure an algorithm will **not exceed** a user-defined **memory constraint**

ON-GOING WORK

Symmetric Matrix Multiplication (SYMM) – application

Our versatile approach is of interest in the **`Fast Methods for Randomized numerical algebra`** (FMR) data mining library.

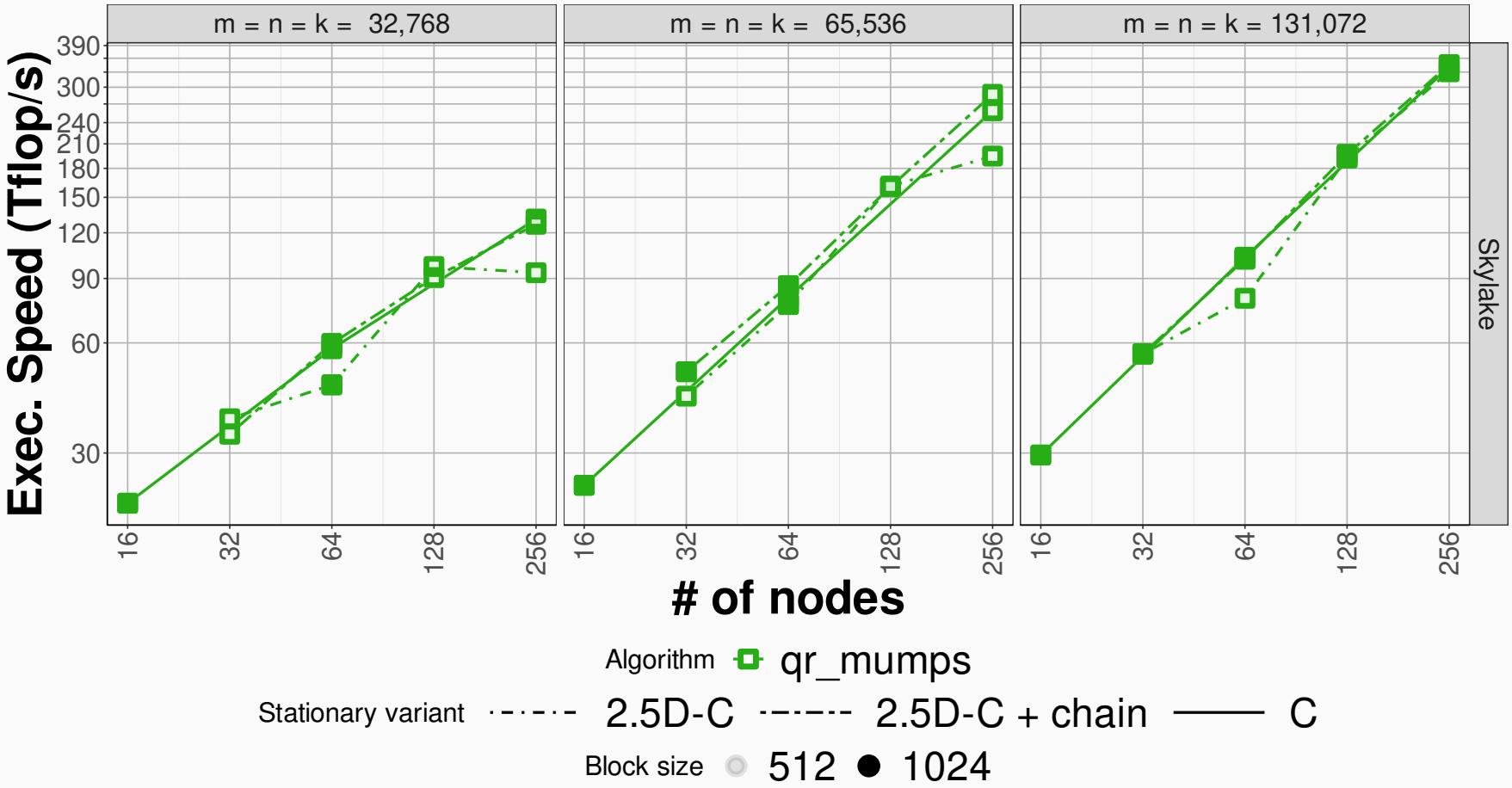


- **A-stationary algorithm**: involved matrices are disproportionate
- **distribution-agnosticism**: to deploy *symmetric distributions*

⇒ In collaboration with:
Olivier Beaumont, Olivier Coulaud, Lionel Eyraus-Dubois, Alain Franc, Romain Peressoni, Florient Pruvost, Mathieu Vérité

Extensive study on communication ordering for GEMM

Experiments questionning the scheduling of communications



⇒ In collaboration with:

Olivier Beaumont, Alexandre Denis, Lionel Eyraus-Dubois, Emmanuel Jeannot, Mathieu Vérité, Philippe Swartvagher

Dense Factorizations (LU, Cholesky)

2.5D factorization algorithms have **more complex** critical path.

They require more advanced key functionalities:

1. Accumulation and reduction of **updates over layers**
2. **Replication** of factorization tasks
3. **All-reduce pattern** to implement tournament pivoting

The programming model currently lacks support for 2. and 3.

2.5D factorization algorithms have **more complex** critical path.

They require more advanced key functionalities:

1. Accumulation and reduction of **updates over layers**
2. **Replication** of factorization tasks
3. **All-reduce pattern** to implement tournament pivoting

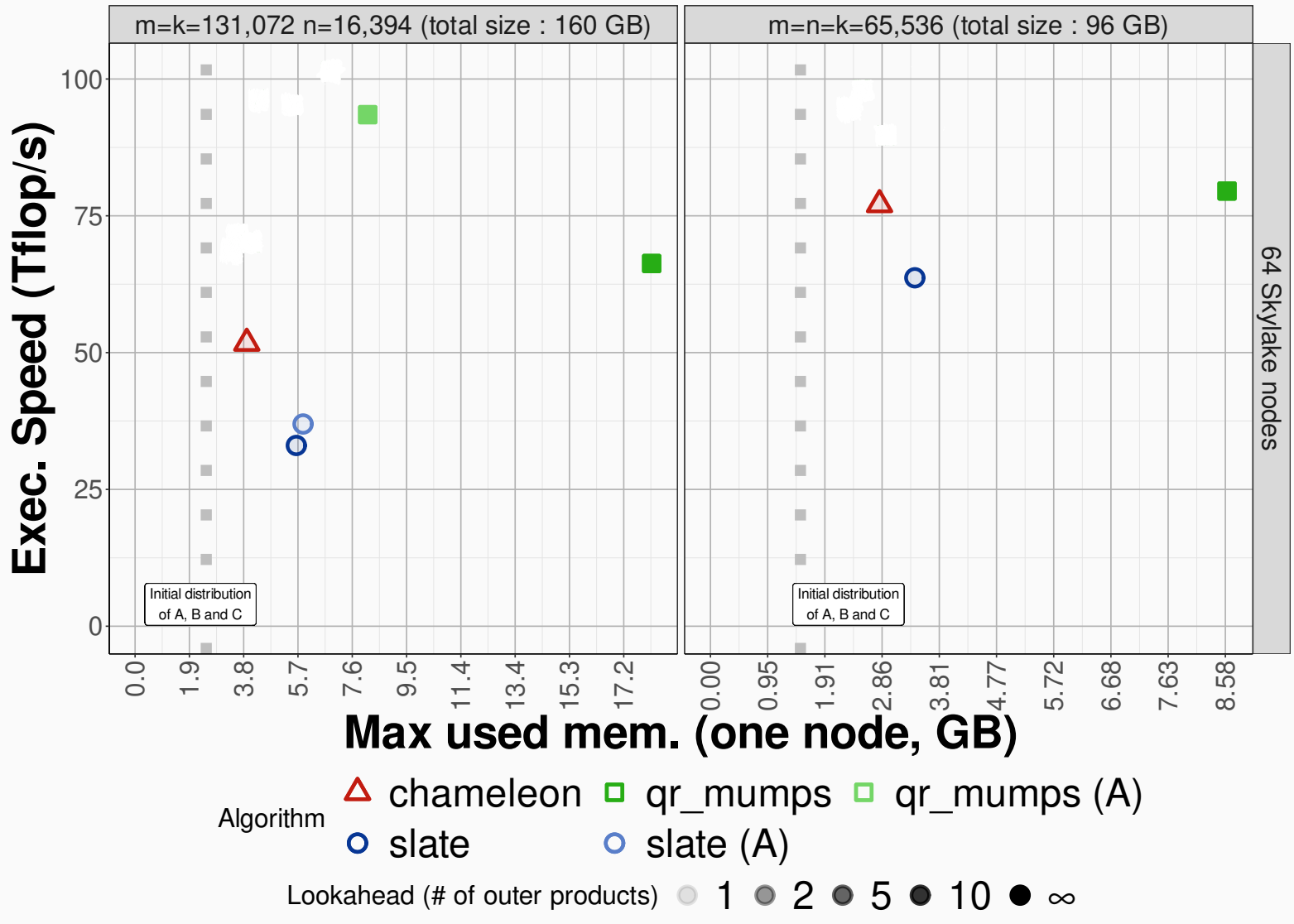
The programming model currently lacks support for 2. and 3.

BACKUP SLIDES

Memory Consumption Concern - experiment

Submitting the **full DAG** allows **greedy** memory consumption

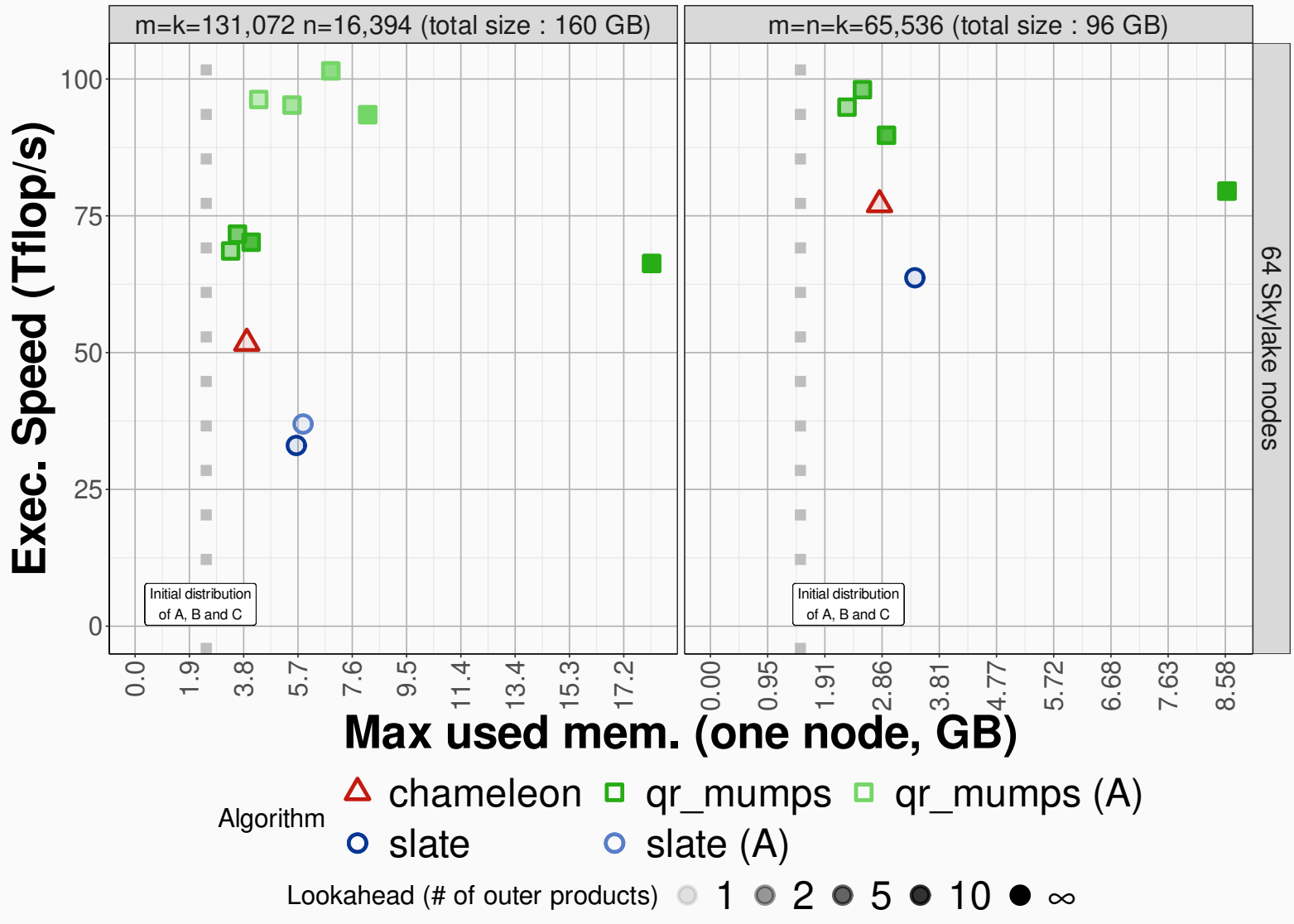
Task window mechanisms can help limit memory consumption



Memory Consumption Concern - experiment

Submitting the **full DAG** allows **greedy** memory consumption

Task window mechanisms can help limit memory consumption



Some common operations are done before switching on the `stat` parameter:

- Binding C to the `init` and `sum` codelets
- (submitting) the scaling of C by β

2.5D-A-Stat

```
do jh=1,n/h; do layer=1,h
  j = jh * layer
  do i=1,m
    do l=1,k
Ail => transa ? A(l,i) : A(i,l)
Blj => transb ? B(j,l) : B(l,j)
Cij => C(i,j)
work_node = Ail%owner
          + r*c*layer
call insert_task( gemm,
  work_node:ON_NODE,
  Ail:R,Blj:R,
  Cij:RANK_REDUX )
    end do
    call mpi_redux_tree(C(i,j))
  end do
end do; end do
```

2.5D-B-Stat

```
do ih=1,m/h; do layer=1,h
  i = ih * layer
  do j=1,n
    do l=1,k
Ail => transa ? A(l,i) : A(i,l)
Blj => transb ? B(j,l) : B(l,j)
Cij => C(i,j)
work_node = Blj%owner
          + r*c*layer
call insert_task( gemm,
  work_node:ON_NODE,
  Ail:R,Blj:R,
  Cij:RANK_REDUX )
    end do
    call mpi_redux_tree(C(i,j))
  end do
end do; end do
```

2.5D-C-Stat

```
do lh=1,k/h; do layer=1,h
  l = lh * layer
  do i=1,m
    do j=1,n
Ail => transa ? A(l,i) : A(i,l)
Blj => transb ? B(j,l) : B(l,j)
Cij => C(i,j)
work_node = Cij%owner
          + r*c*layer
call insert_task( gemm,
  work_node:ON_NODE,
  Ail:R,Blj:R,
  Cij:RANK_REDUX )
    end do
  end do
end do; end do
do i=1,m; do j=1,n
  call mpi_redux_tree(C(i,j))
end do; end do
```

StarPU – additional features ?

Some additional features to StarPU have been considered.

- Automatically insert reduction tasks
  - Further simplify factorization implementations

```
call starpu_mpi_redux(A_{I,I})
call insert_task(potrf, A_{I,I}:RW)
```

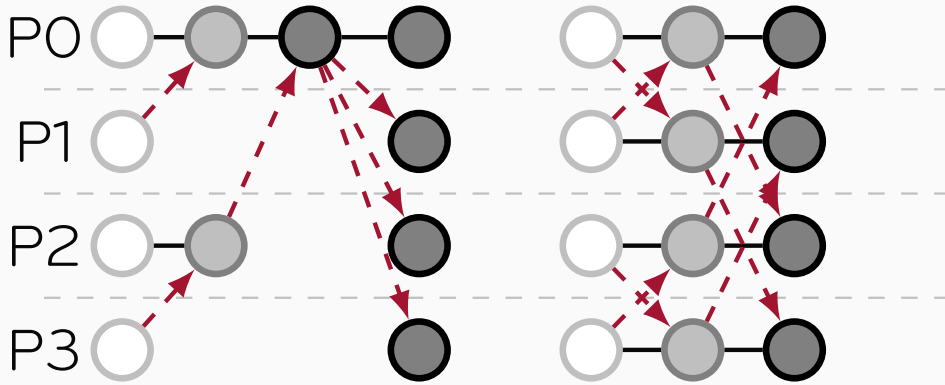
⇓

```
! implicit reduction
call insert_task(potrf, A_{I,I}:RW)
```

- Allow tasks replication across nodes
  - Required by state-of-the-art algorithms
  - Each resulting tile can be sent to different subsets of nodes

```
call insert_task(potrf, A_{I,I}:RW,
                 ON_NODES:task_map[I,I,:])
call insert_task(trsm,  A_{I,I}:R,  A_{i,I}:RW,
                 FROM_NODE:task_map[I,I,i%h])
```

- Allow all-reduce pattern
  - Shorten the critical path as replicated data is already available



Ideal expression - task mappings

- 2D-C-Stat owner of C computes
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RW`

- 2D-A-Stat owners of A compute
  - `task_map[i,j,l]` $= l \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RANK_REDUX`

- 2.5D-C-Stat several nodes compute
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r) + rc * \frac{l}{k/h}$
  - `access_mode` $=$ `RANK_REDUX`

The advanced model allows to **cover all variants** of SUMMA algorithms with three nested loops.
In a **distribution-agnostic** fashion.

Ideal expression - task mappings

- 2D-C-Stat owner of C computes
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RW.and.COMMUTE`

- 2D-A-Stat owners of A compute
  - `task_map[i,j,l]` $= l \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RANK_REDUX`

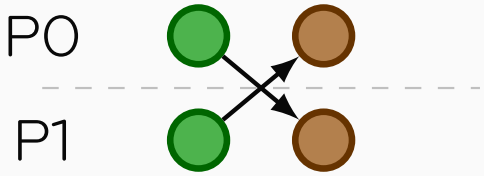
- 2.5D-C-Stat several nodes compute
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r) + rc * \frac{l}{k/h}$
  - `access_mode` $=$ `RANK_REDUX`

The advanced model allows to **cover all variants** of SUMMA algorithms with three nested loops.
In a **distribution-agnostic** fashion.

Ideal expression - task mappings

- 2D-C-Stat owner of C computes
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RW.and.COMMUTE`

- 2D-A-Stat owners of A compute
  - `task_map[i,j,l]` $= l \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RANK_REDUX`



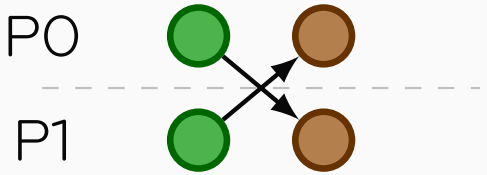
- 2.5D-C-Stat several nodes compute
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r) + rc * \frac{l}{k/h}$
  - `access_mode` $=$ `RANK_REDUX`

The advanced model allows to **cover all variants** of SUMMA algorithms with three nested loops.
In a **distribution-agnostic** fashion.

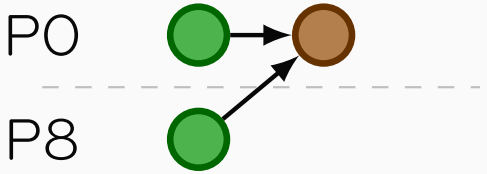
Ideal expression - task mappings

- 2D-C-Stat owner of C computes
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RW.and.COMMUTE`

- 2D-A-Stat owners of A compute
  - `task_map[i,j,l]` $= l \,\%\, c + c * (i \,\%\, r)$
  - `access_mode` $=$ `RANK_REDUX`



- 2.5D-C-Stat several nodes compute
  - `task_map[i,j,l]` $= j \,\%\, c + c * (i \,\%\, r) + rc * \frac{l}{k/h}$
  - `access_mode` $=$ `RANK_REDUX`

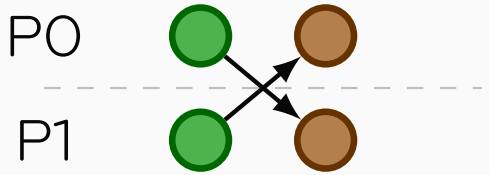


The advanced model allows to **cover all variants** of SUMMA algorithms with three nested loops.
In a **distribution-agnostic** fashion.

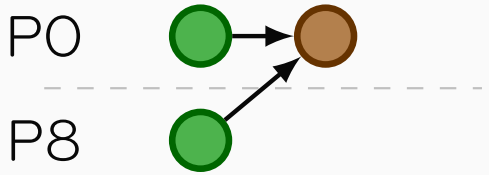
Ideal expression - task mappings

- 2D-C-Stat owner of C computes
  - `task_map[i,j,l]` = $C_{i,j}.owner\_rank$
  - `access_mode` = `RW.and.COMMUTE`

- 2D-A-Stat owners of A compute
  - `task_map[i,j,l]` = $A_{i,l}.owner\_rank$
  - `access_mode` = `RANK_REDUX`



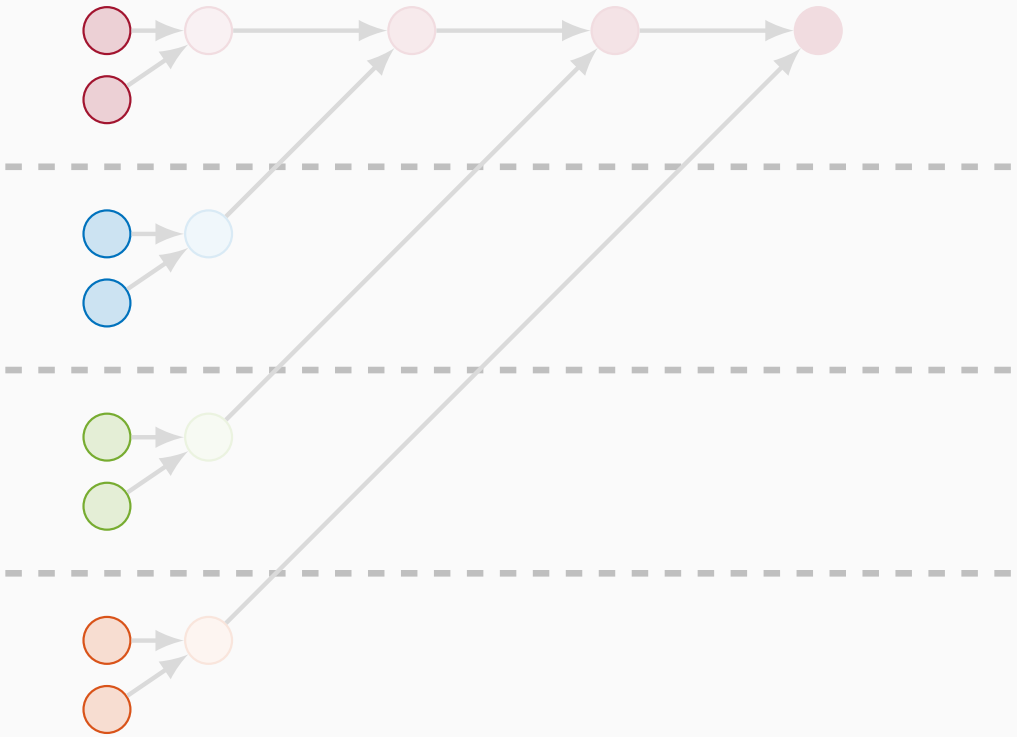
- 2.5D-C-Stat several nodes compute
  - `task_map[i,j,l]` = $C_{i,j}.owner\_rank + rc * \frac{l}{k/h}$
  - `access_mode` = `RANK_REDUX`



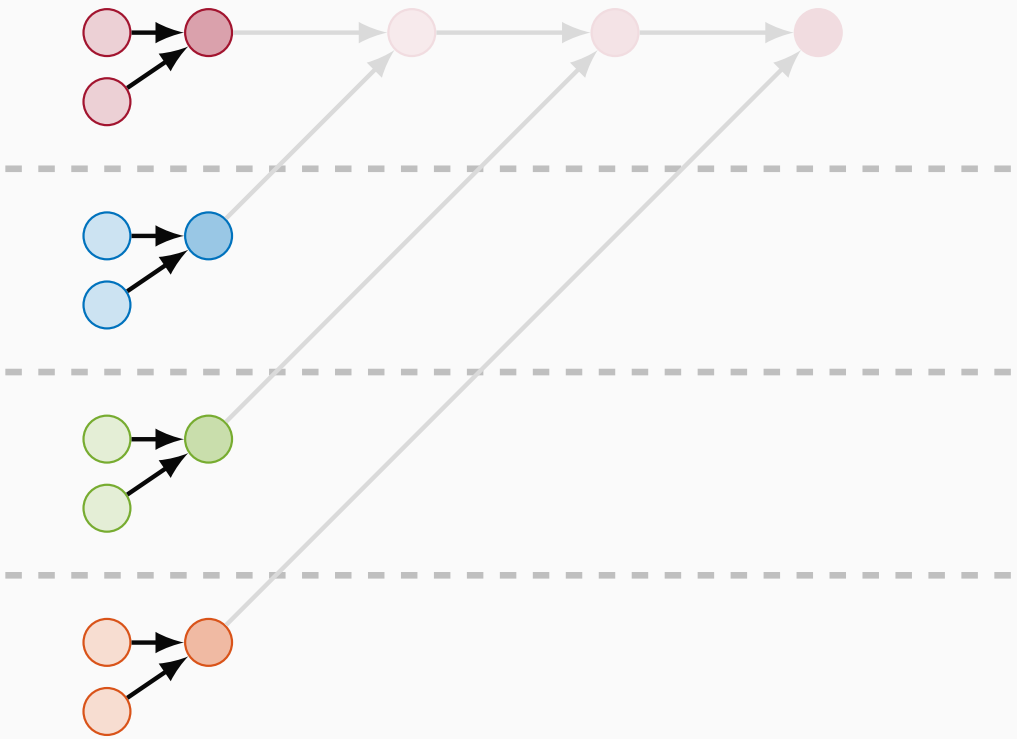
The advanced model allows to **cover all variants** of SUMMA algorithms with three nested loops.
In a **distribution-agnostic** fashion.

Features of StarPU – 1.3

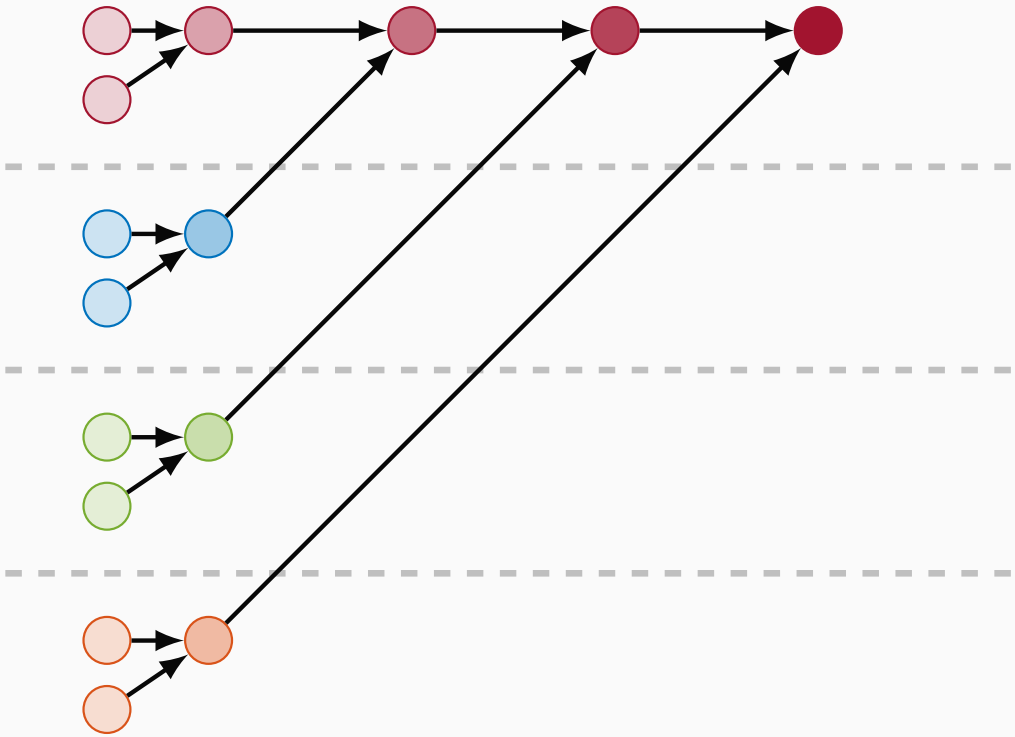
- `STARPU_REDUX`
  - One contribution per **core**
    - Intra-node reduction through binary tree
  - Remain available in StarPU (master)
- `starpu_mpi_redux_data`
  - **All** nodes in `MPI_COMM_World` contribute through a **flat** tree
  - **Blocking** method



- `STARPU_REDUX`
  - One contribution per **core**
    - Intra-node reduction through binary tree
  - Remain available in StarPU (master)
- `starpu_mpi_redux_data`
  - **All** nodes in `MPI_COMM_World` contribute through a **flat** tree
  - **Blocking** method

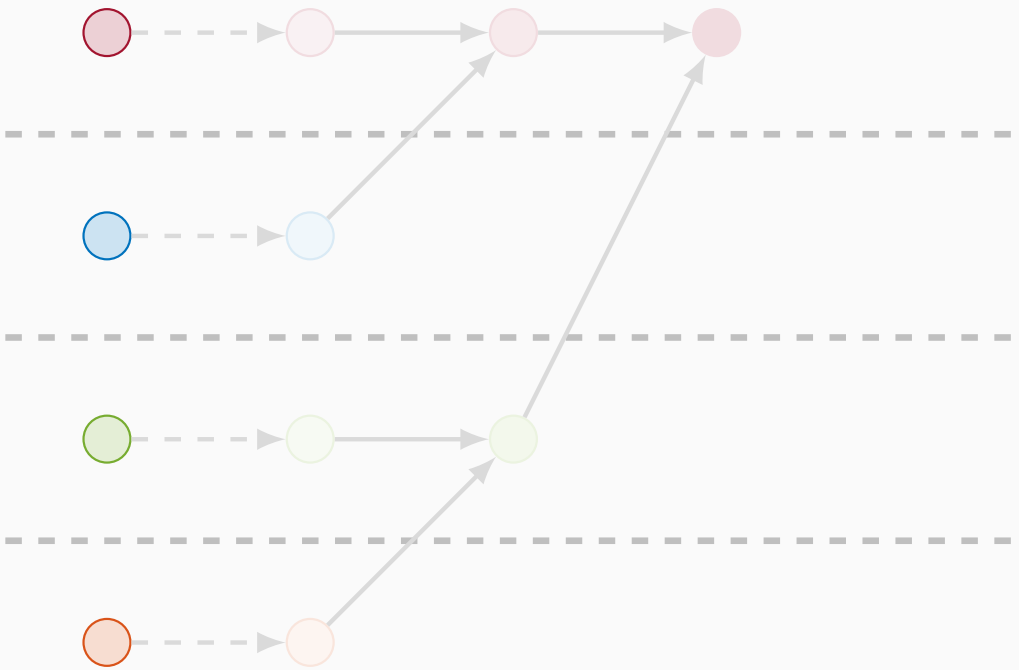


- `STARPU_REDUX`
  - One contribution per **core**
    - Intra-node reduction through binary tree
  - Remain available in StarPU (master)
- `starpu_mpi_redux_data`
  - **All** nodes in `MPI_COMM_World` contribute through a **flat** tree
  - **Blocking** method



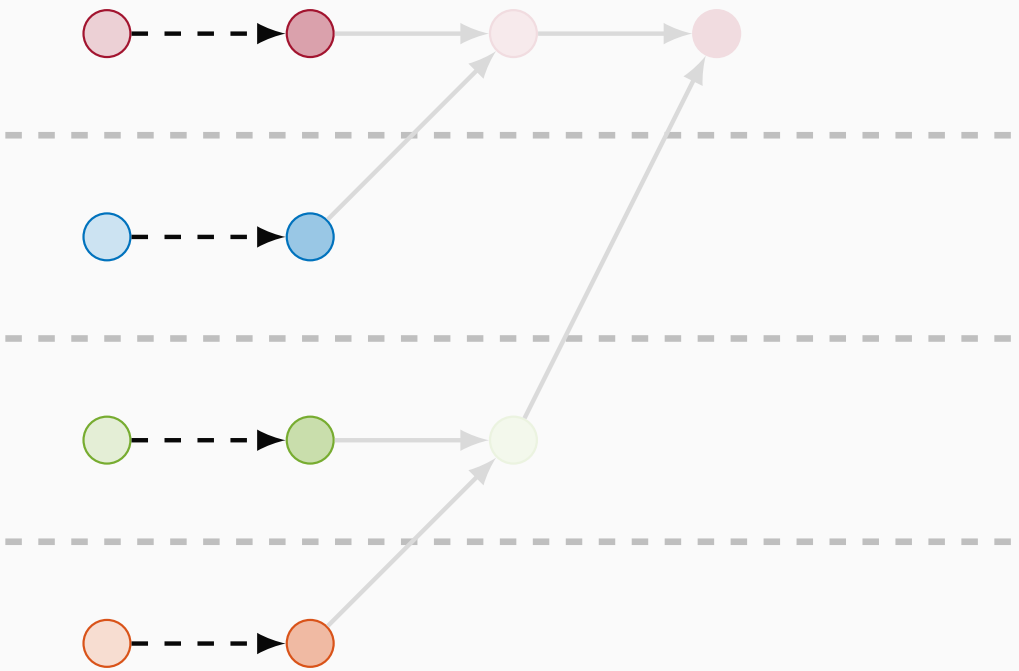
Features of StarPU – our contribution

- `STARPU_MPI_REDUX` (**Feature!** issue 3)
  - One contribution per **contributing node**
  - Interpreted as `STARPU_RW|STARPU_COMMUTE`
- `starpu_mpi_redux_data_tree` (**Enhancement!** merge req. 21)
  - All **contributing** nodes join in the reduction through a **n-ary** tree
  - **Non-blocking** method



Features of StarPU – our contribution

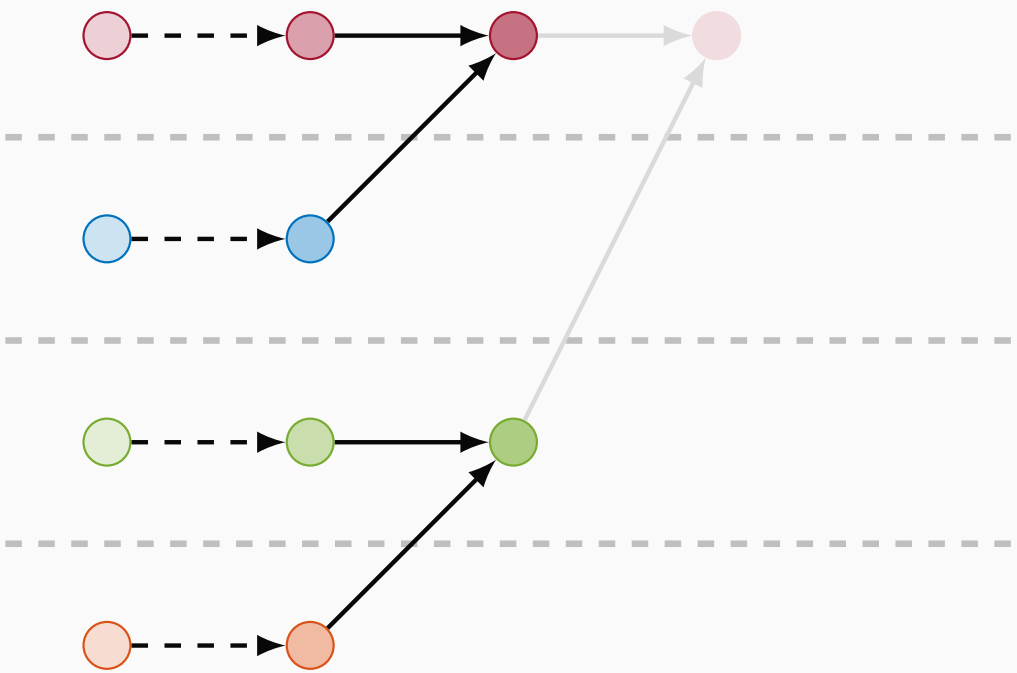
- `STARPU_MPI_REDUX` (**Feature!** issue 3)
  - One contribution per **contributing node**
  - Interpreted as `STARPU_RW|STARPU_COMMUTE`
- `starpu_mpi_redux_data_tree` (**Enhancement!** merge req. 21)
  - All **contributing** nodes join in the reduction through a **n-ary** tree
  - **Non-blocking** method



X

Features of StarPU – our contribution

- `STARPU_MPI_REDUX` (**Feature!** issue 3)
  - One contribution per **contributing node**
  - Interpreted as `STARPU_RW|STARPU_COMMUTE`
- `starpu_mpi_redux_data_tree` (**Enhancement!** merge req. 21)
  - All **contributing** nodes join in the reduction through a **n-ary** tree
  - **Non-blocking** method



Features of StarPU – our contribution

- `STARPU_MPI_REDUX` (**Feature!** issue 3)
  - One contribution per **contributing node**
  - Interpreted as `STARPU_RW|STARPU_COMMUTE`
- `starpu_mpi_redux_data_tree` (**Enhancement!** merge req. 21)
  - All **contributing** nodes join in the reduction through a **n-ary** tree
  - **Non-blocking** method