FUTURE OF MPI RMA: IF IT’S NOT BAROQUE DON’T FIX IT

TORSTEN HOEFLER, KEITH UNDERWOOD, JEFF HAMMOND, AND BILL GROPP

MODERATOR: JAMES DINAN

INTRODUCING OUR PANELISTS

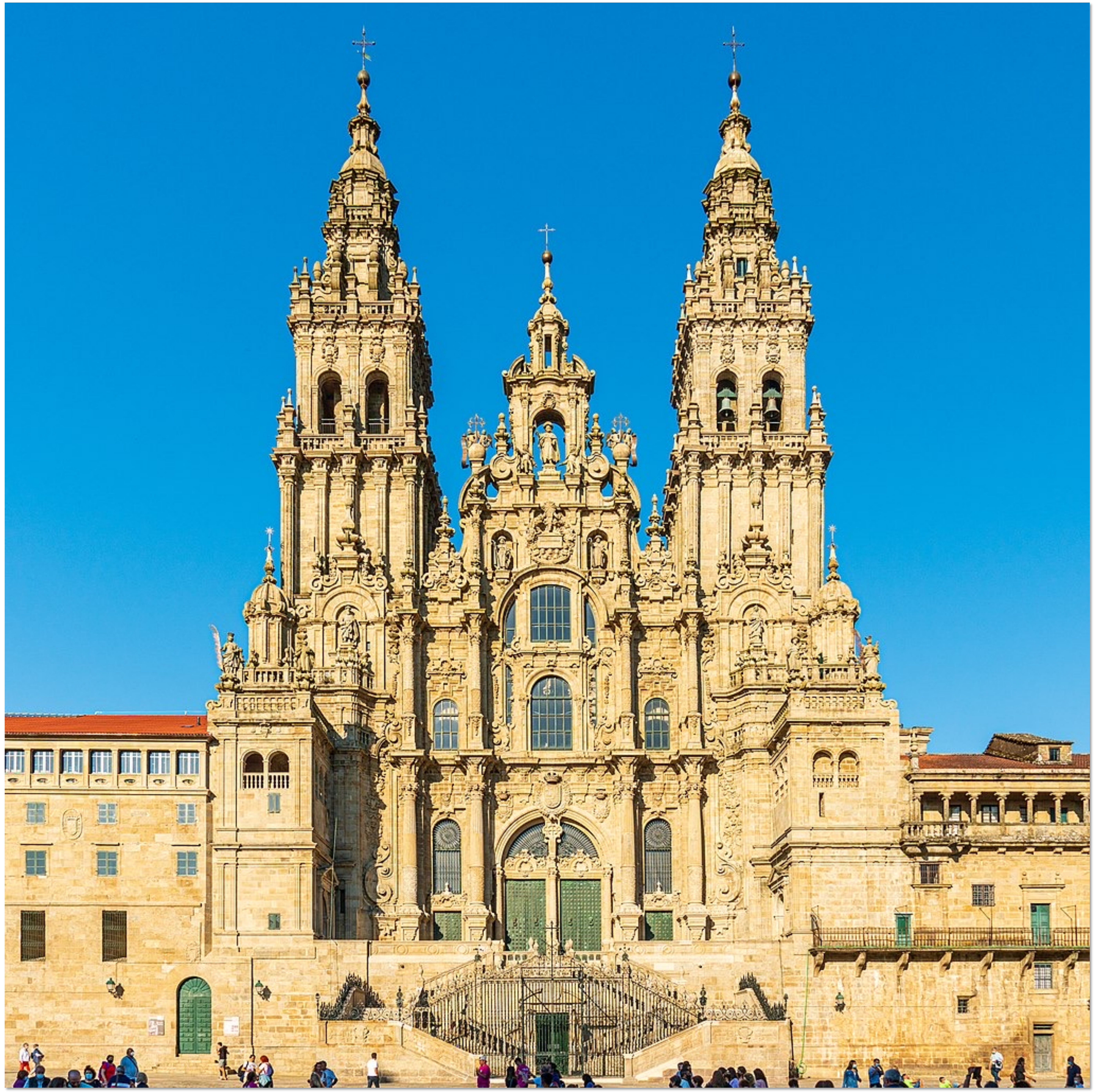
Jeff Hammond, Bill Gropp, Torsten Hoefler, and Keith Underwood

BAROQUE

Highly ornate and extravagant in style

Credit: Disney's Beauty and the Beast

SANTIAGO DE COMPOSTELA ARCHCATHEDRAL BASILICA. GALICIA, SPAIN



By Fernando - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=108259683

SANTIAGO DE COMPOSTELA ARCHCATHEDRAL BASILICA. GALICIA, SPAIN

Separate Memory Model
Unified Memory Model
Active Target Synchronization
Passive Target Synchronization
MPI_Win_allocate Flavor
MPI_Win_create Flavor
MPI_Win_create_dynamic Flavor
MPI_Win_allocate_shared Flavor
Store
Accumulate Ops
Load
Put
Get

By Fernando - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=108259683

SANTIAGO DE COMPOSTELA ARCHCATHEDRAL BASILICA. GALICIA, SPAIN

Separate Memory Model

Unified Memory Model

Steward of NOSUCCEED

Window lock

Active Target Synchronization

Passive Target Synchronization

MODE_NOCHECK Incantation

Patron saint of post, start, complete, wait

Turned to stone for insisting put is local

Guardian of overlapping windows

MPI_Win_allocate Flavor

MPI_Win_create Flavor

MPI_Win_create_dynamic Flavor

MPI_Win_allocate_shared Flavor

Store

Accumulate Ops

Load

Put

MIN/MAXLOC

Get

MPI_AINT yet hit MPI_ROCK_BOTTOM

Noncontig. MPI Datatype

Accumulate Ordering

Displacement Unit

help

MPI RMA User

By Fernando - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=108259683

PANELISTS, PLEASE HELP US SORT OUT …

1. What usage models should drive RMA?

2. What aspects of system architecture will drive the future of MPI RMA?
   - How can we strike the right balance between portability and performance?

3. What new “killer features” should we add?

4. Do we start from scratch or from MPI 4.0?

Panelists’ Slides

MPI-RMA history and future

The good (very) old MPI-2 days (1997)

- Very elegant interface focused on algorithms
- Designed for message-centric hardware
  - Not what we had ten years later

The good (also) old MPI-3 days (2012)

- Observation: suboptimal performance of MPI-2 RMA
- Adopted to hardware at the time™
  - RDMA transports
  - Added limited atomics

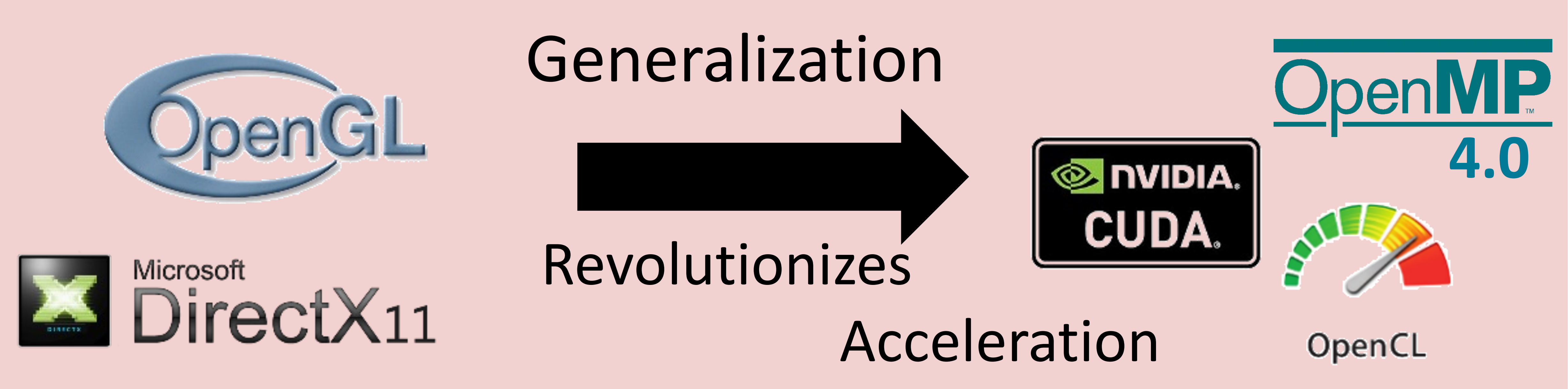
Another 10 years later (today!)

- Observation: suboptimal performance for some tasks
  - Bill Gropp: “sequence of atomics” and many more
- Are we back at square one?
  - Repeat: “It was a good idea at the time”™

The next era: Smart NICs and In-network Compute!

- Smart NICs are ubiquitous – simple computations
  - RMA is a baby step towards network acceleration
- Needs an MPI-like standard specification
  - sPIN is one possible proposal

Established Principles for Compute Acceleration

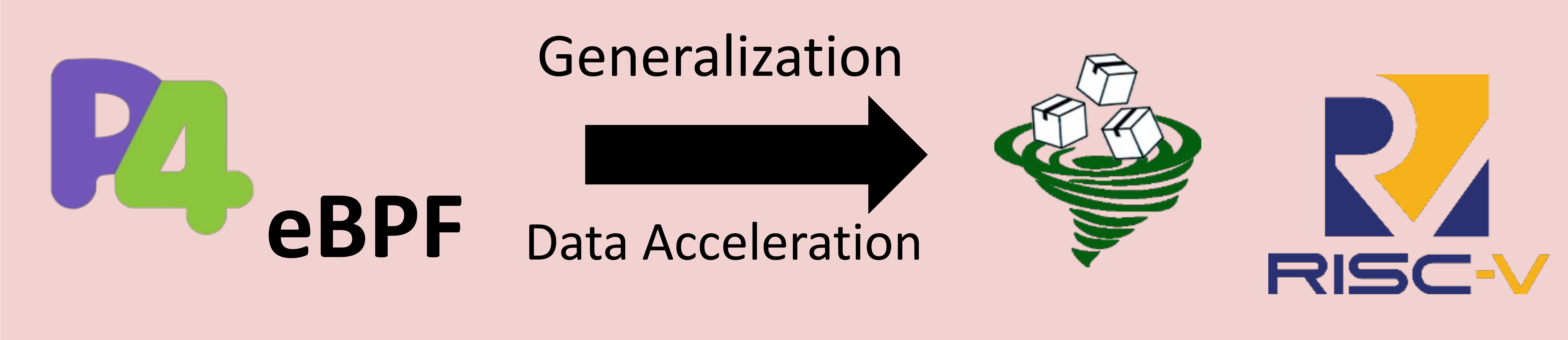


Generalization

Revolutionizes

Acceleration

Where do we stand in Network Acceleration?



eBPF

Generalization

Data Acceleration

MPI-RMA history and future – find all the details online on YouTube!

The good (also) old MPI-3 days (2012)

- Observation: suboptimal performance of MPI-2 RMA

YouTube CH https://www.youtube.com/watch?v=t6jdjnnIRZs

ETH zürich DINFK

T. Hoefler

with S. Di Girolamo, K. Taranov, D. De Sensi, L. Benini, R. E. Grant, R. Brightwell, A. Kurth, M. Schaffner, T. Schneider, J. Beránek, M. Besta, L. Be

General in-network processing – time is ripe!

Keynote at the High-Performance Interconnects Forum with HPC China 2020

General in-network processing - time is ripe!

450 views • Oct 1, 2020

YouTube CH https://www.youtube.com/watch?v=Jfn8LusAR1I

ETH zürich DINFK

T. Hoefler

with S. Di Girolamo, D. De Sensi, K. Taranov, L. Benini, R. E. Grant, R. Brightwell, A. Kurth, M. Schaffner, T. Schneider, J. Beránek, M. Besta, D. Roweth

New trends for sPIN-based in-network computing: from sparse reductions to RISC-V acceleration!

New trends for sPIN-based in-network computing - from sparse reductions to RISC-V

263 views • Nov 24, 2021

Microsoft DirectX11 Revolutionizes NVIDIA CUDA 4.0 OpenCL Acceleration

Successes and Failures of MPI-3 RMA

Motivation: offer a path to bring “PGAS” capabilities to more users

- Outcome: OpenSHMEM became an active standard development activity

Success: added more “SHMEM-like” RMA capabilities

- Failure: Interface was still too complicated to attract SHMEM users

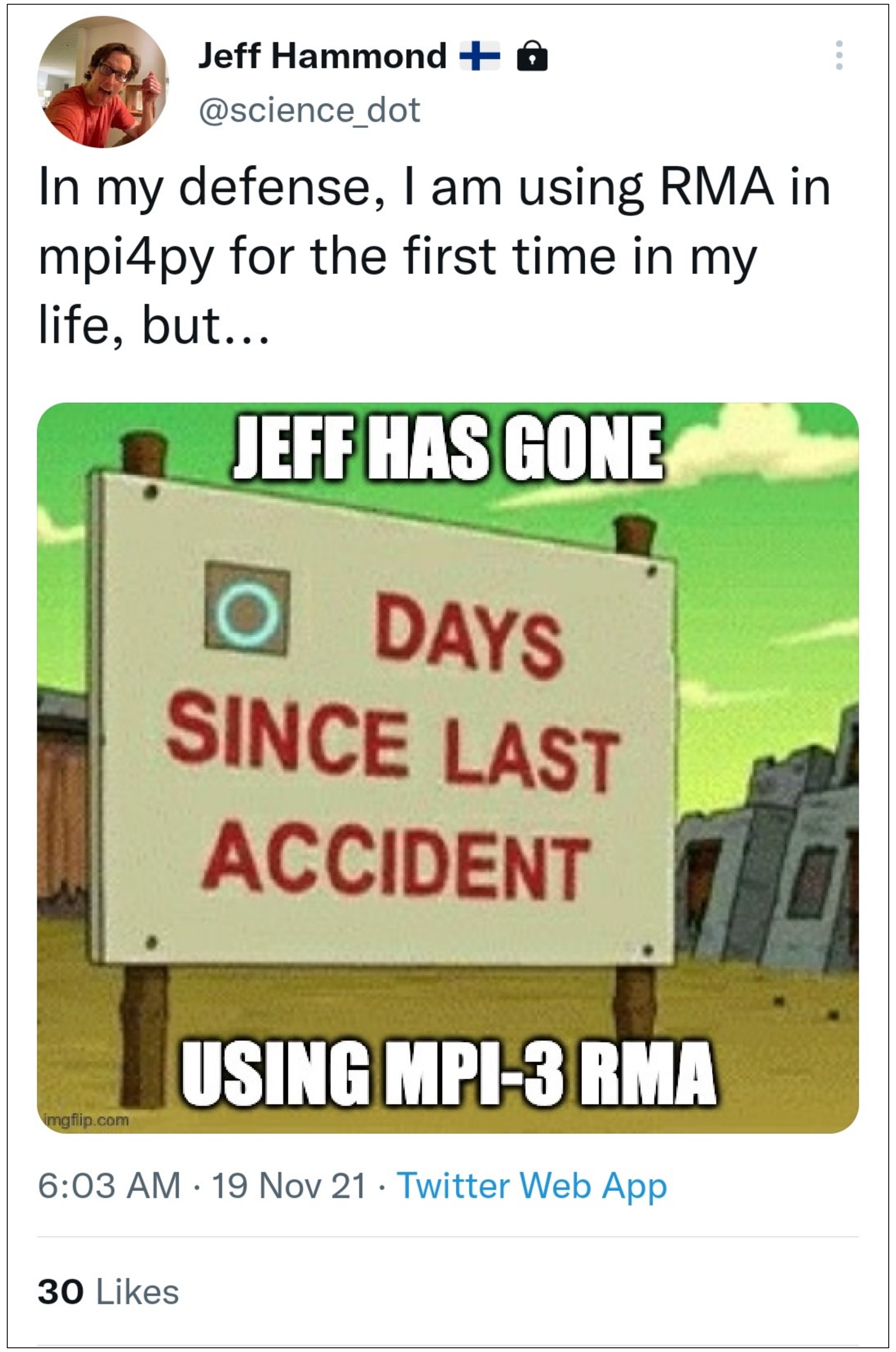
Success: new capabilities could exploit “fast PGAS hardware”

- Failure: But nobody did enough of the implementation work
  - Because there weren’t any customers using it
  - Nobody wanted to solve the chicken and egg problem

A Modest Proposal

Questions for the Panel

WHAT USAGE MODELS SHOULD DRIVE RMA?



Jeff Hammond 🇫🇮 🔒
@science_dot

In my defense, I am using RMA in mpi4py for the first time in my life, but…

6:03 AM · 19 Nov 21 · Twitter Web App

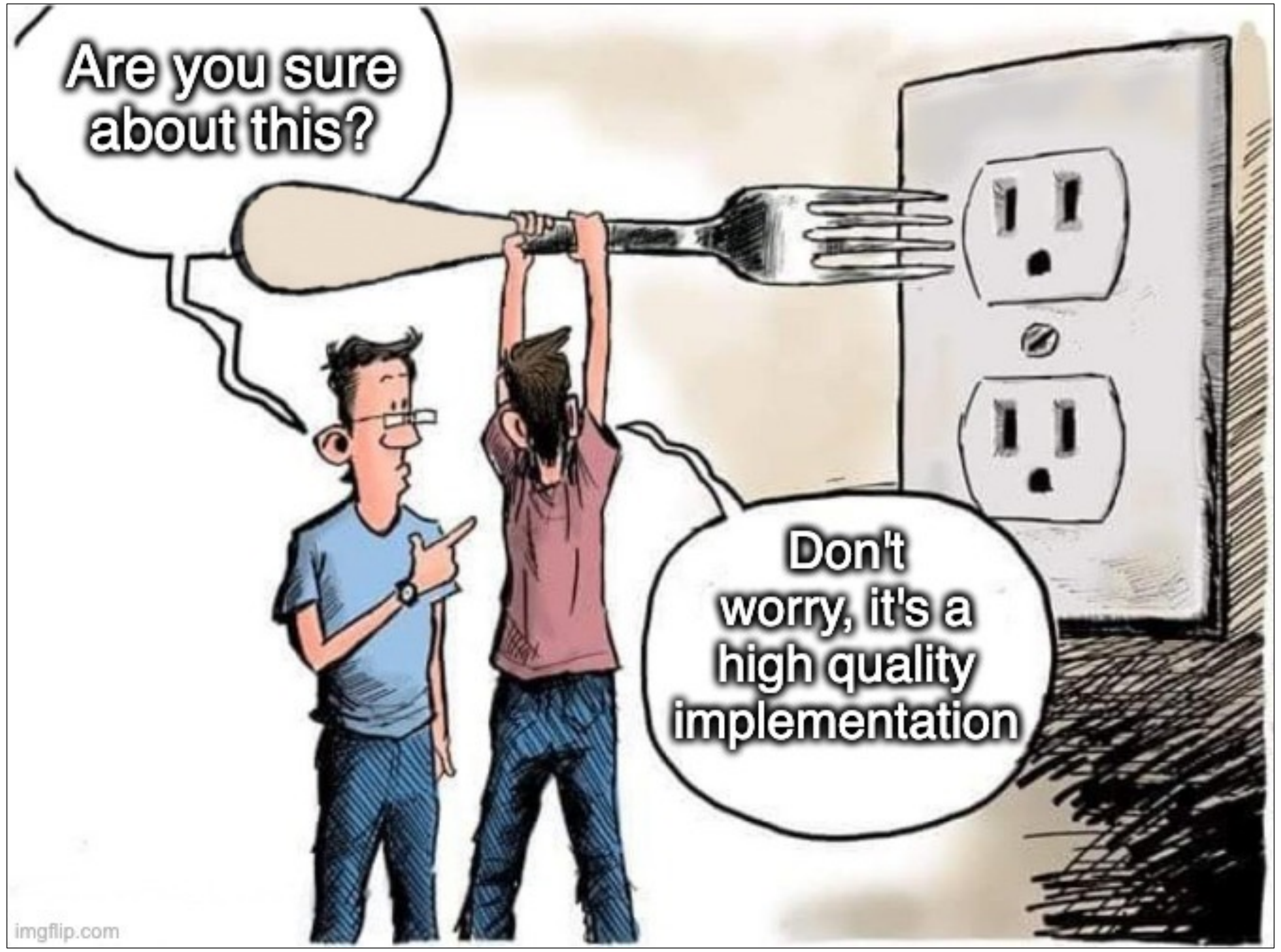
30 Likes

WHAT ASPECTS OF SYSTEM ARCHITECTURE WILL DRIVE MPI RMA?

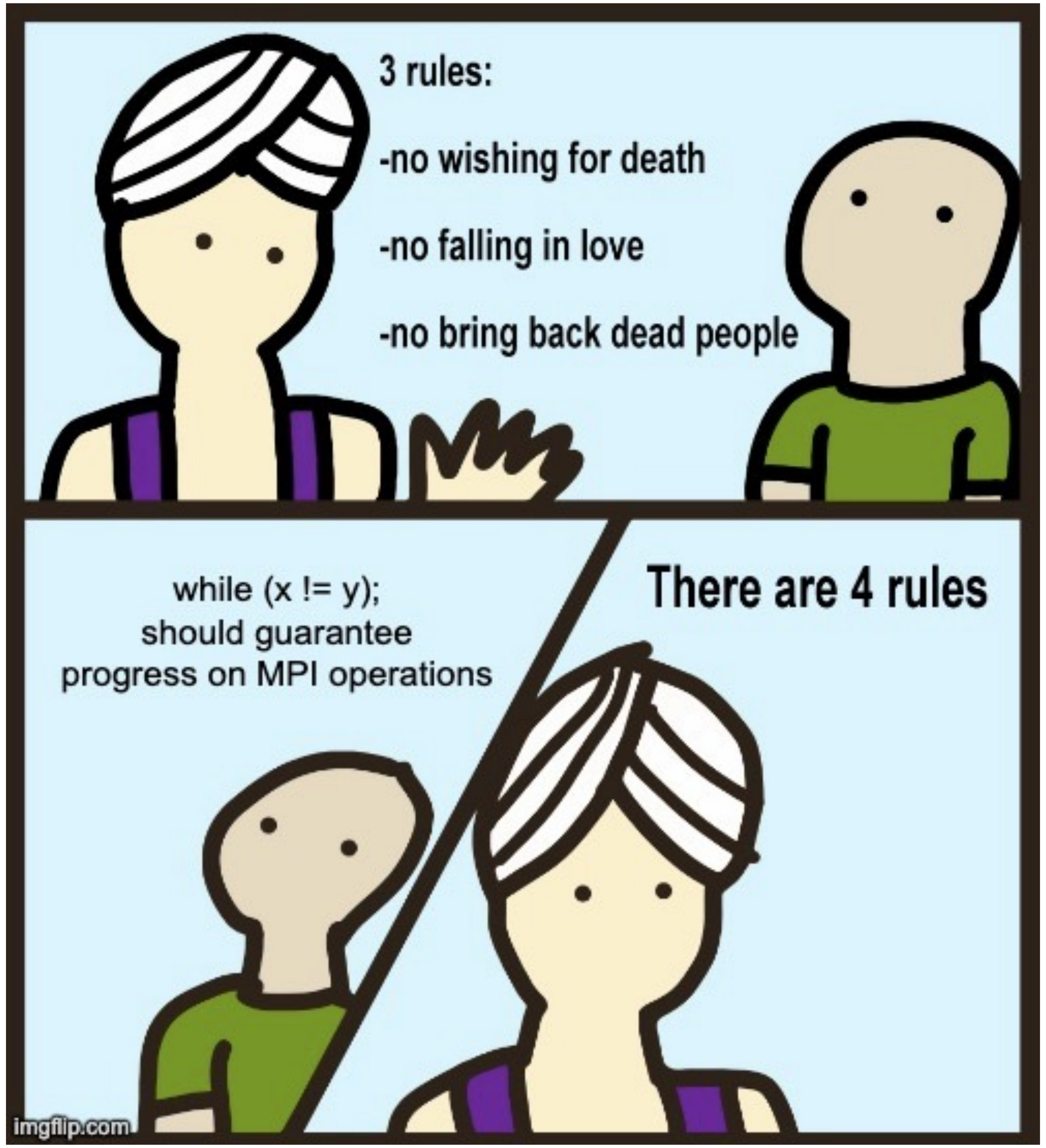
And why is it accelerator-initiated communication?

PROGRAMMABLE
YOUR NETWORK WILL BE
imgflip.com

HOW DO WE STRIKE THE RIGHT BALANCE BETWEEN ENABLING PERFORMANCE AND PROVIDING PORTABILITY?

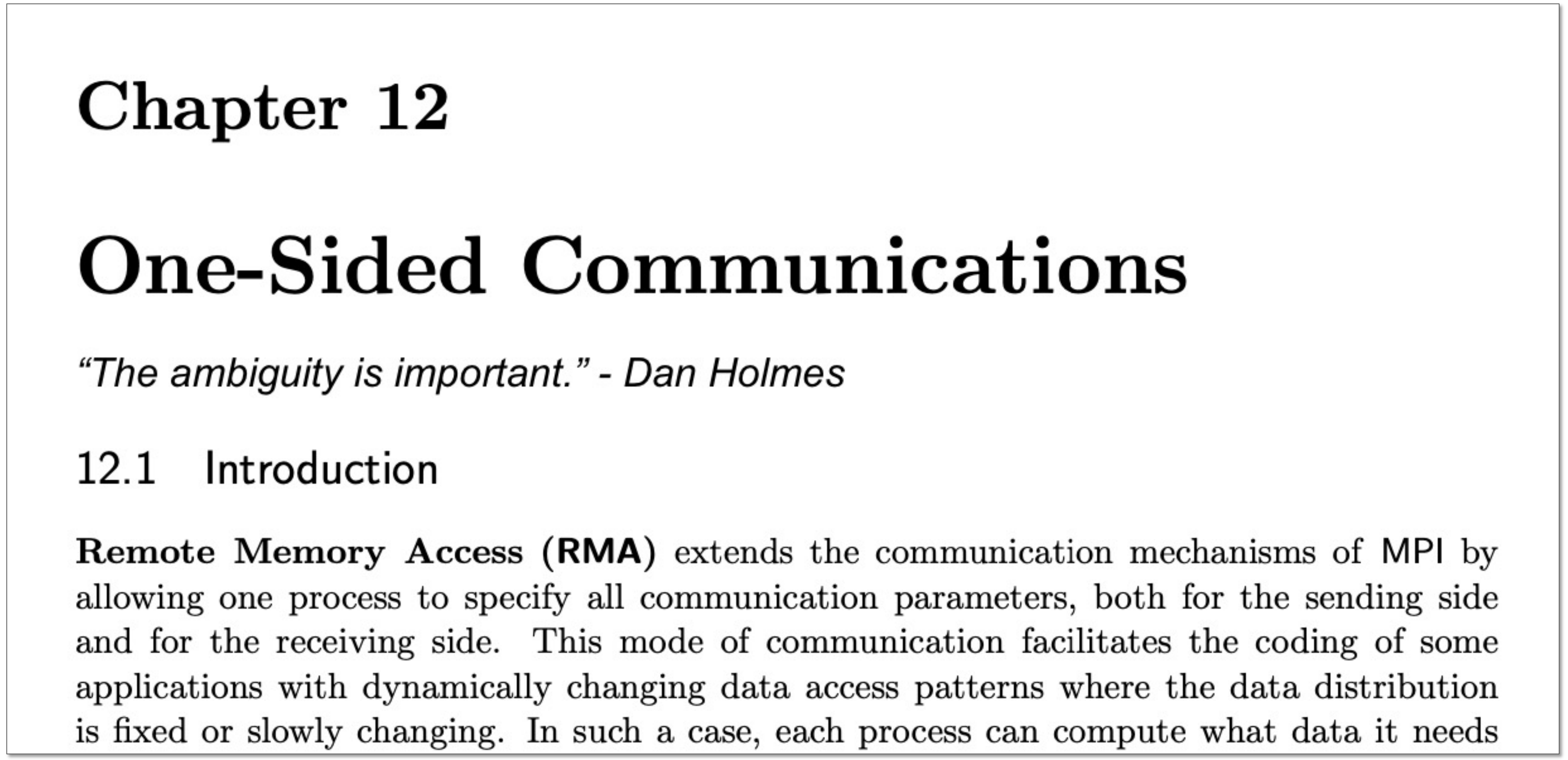


WHAT NEW FEATURES ARE NEEDED IN RMA?



DO WE START OVER FROM SCRATCH?

~Baroque → ~Fix, Baroque → Fix ?



Chapter 12

One-Sided Communications

“The ambiguity is important.” - Dan Holmes

12.1 Introduction

Remote Memory Access (RMA) extends the communication mechanisms of MPI by allowing one process to specify all communication parameters, both for the sending side and for the receiving side. This mode of communication facilitates the coding of some applications with dynamically changing data access patterns where the data distribution is fixed or slowly changing. In such a case, each process can compute what data it needs

Thank You!

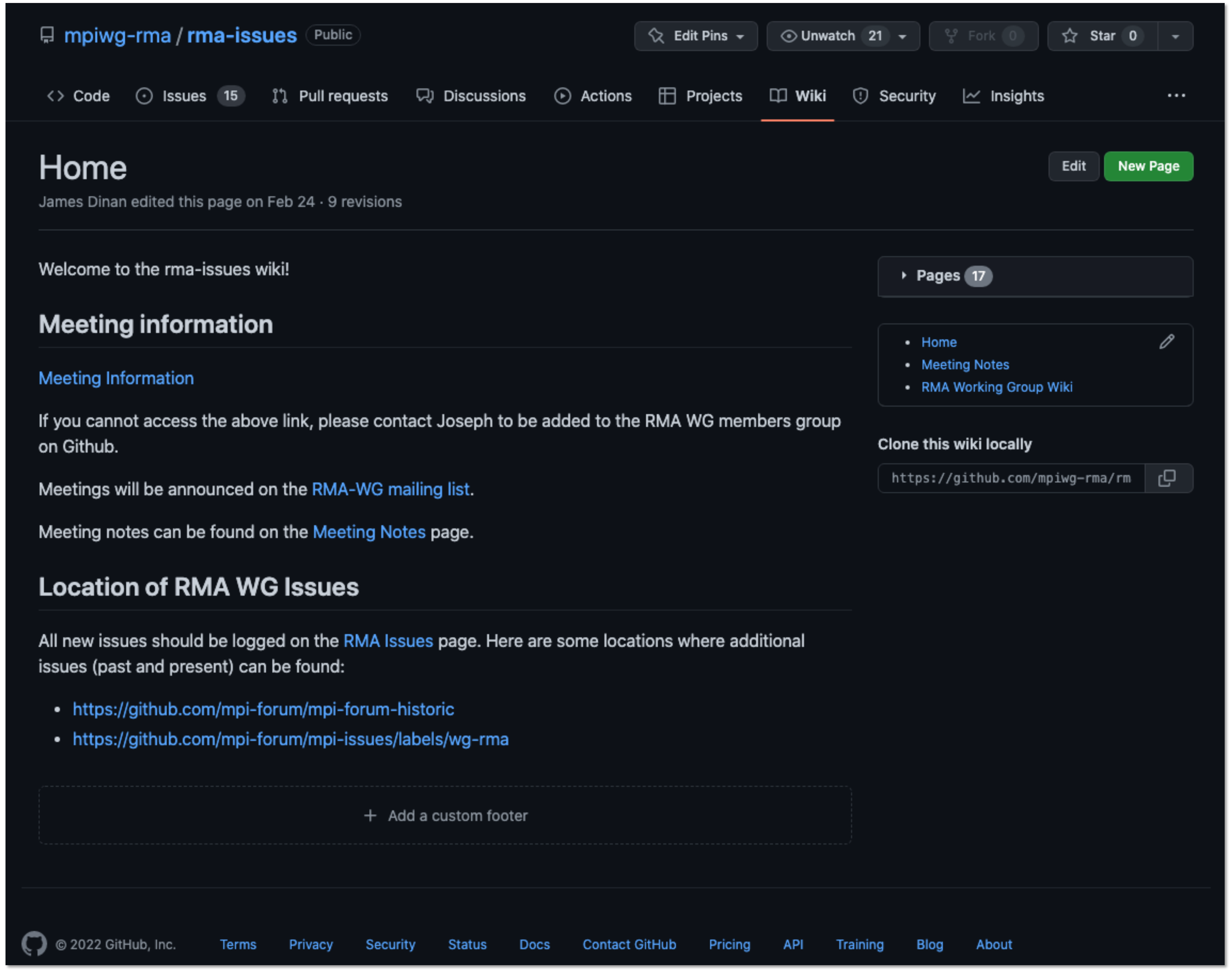
- Panelists:
    - Torsten Hoefler
    - Keith Underwood
    - Jeff Hammond
    - Bill Gropp (Keynote Speaker)
- Workshop organizers:
    - Joseph Schuchart
    - Bill Gropp
    - Jim Dinan
- Speakers and attendees

The Discussion Continues …

MPI RMA Working Group:

- Biweekly meetings
- Thursdays 10:00 am – 11:00pm CT

https://github.com/mpiwg-rma/rma-issues



Home

James Dinan edited this page on Feb 24 · 9 revisions

Welcome to the rma-issues wiki!

Meeting information

Meeting Information

If you cannot access the above link, please contact Joseph to be added to the RMA WG members group on Github.

Meetings will be announced on the RMA-WG mailing list.

Meeting notes can be found on the Meeting Notes page.

Location of RMA WG Issues

All new issues should be logged on the RMA Issues page. Here are some locations where additional issues (past and present) can be found:

- https://github.com/mpi-forum/mpi-forum-historic
- https://github.com/mpi-forum/mpi-issues/labels/wg-rma

NVIDIA®