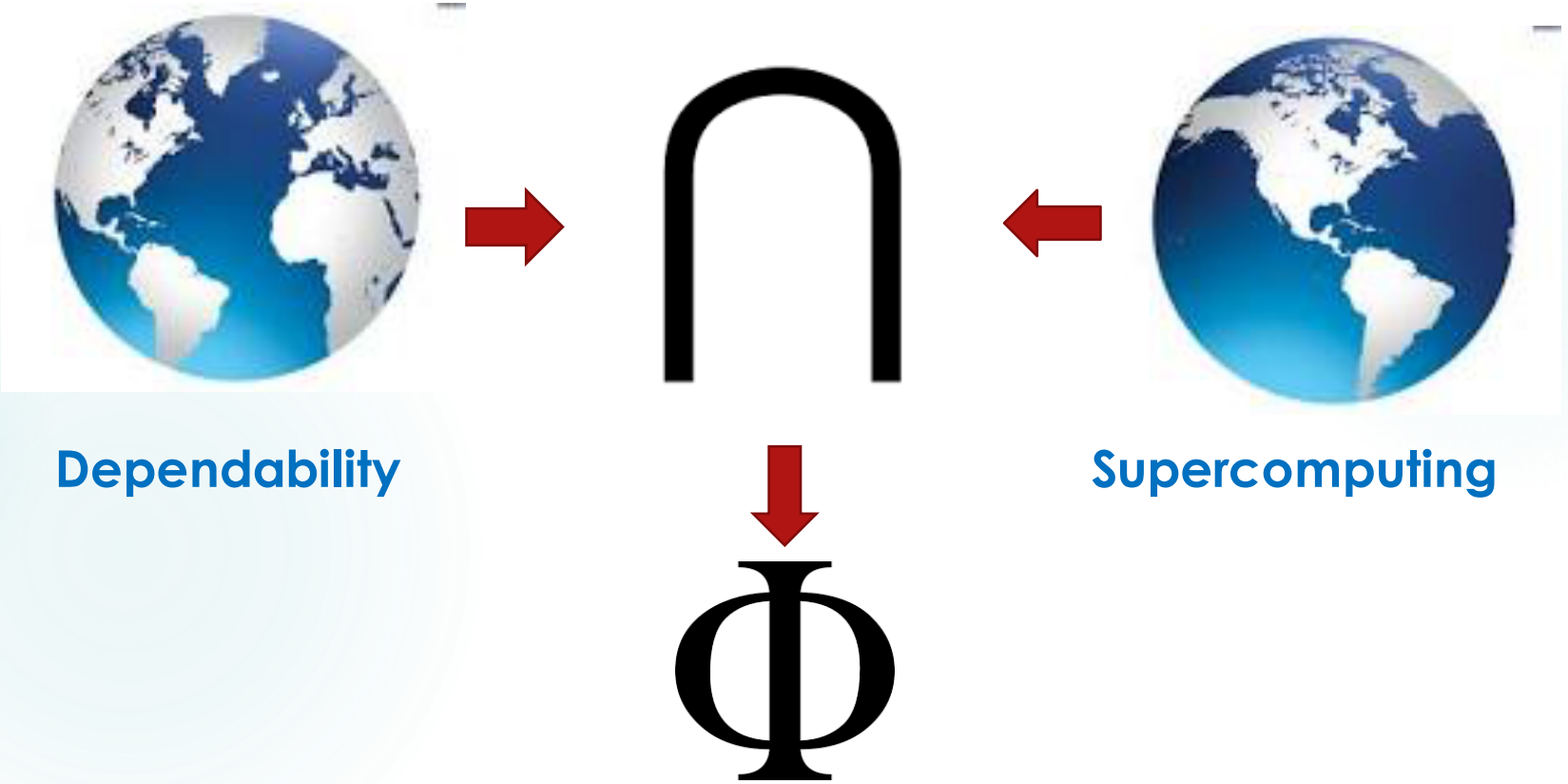
An Overview of Resilience in MPI 3.0

MOOTAZ ELNOZAHY

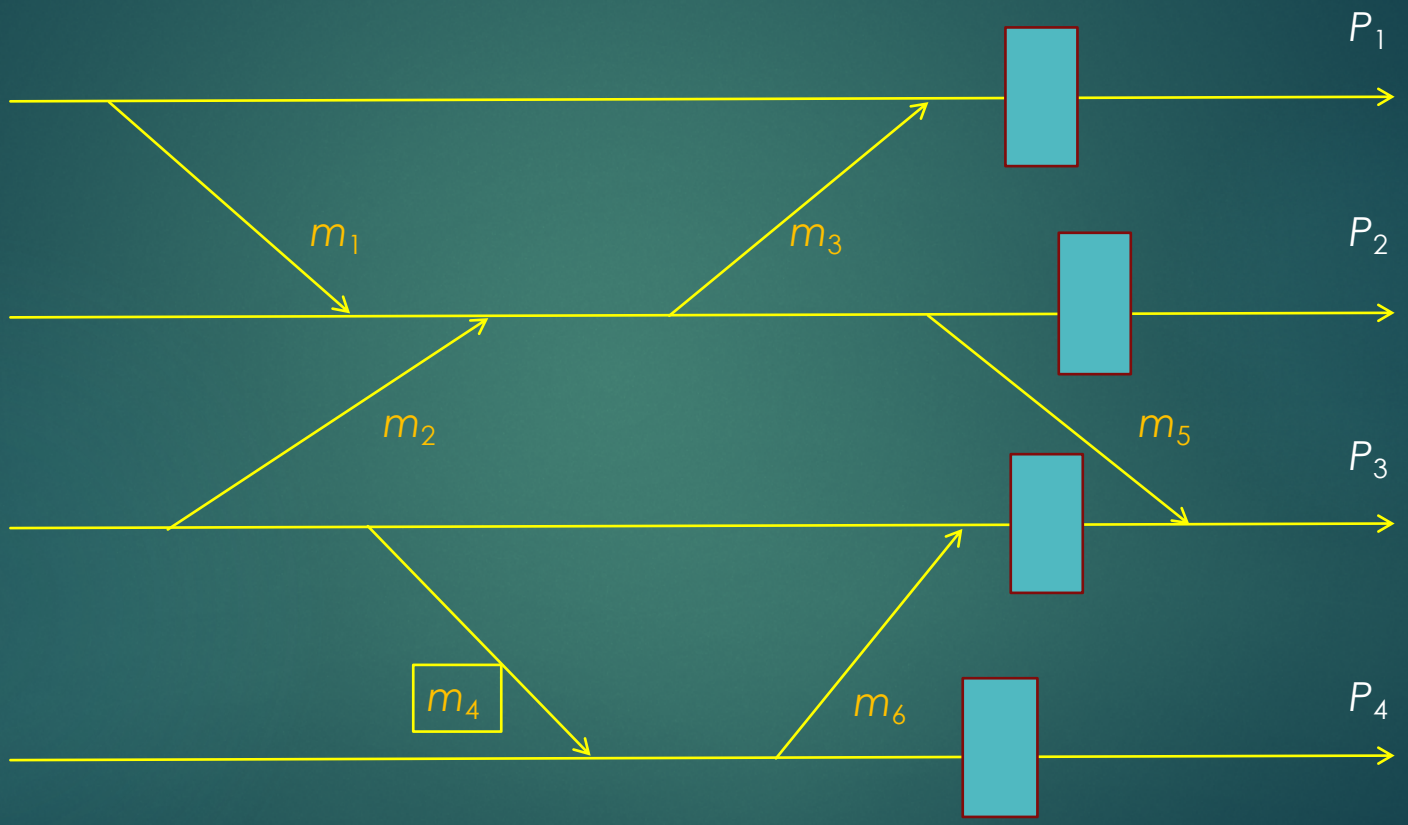
KAUST

Why this presentation?



A comparison of how two communities approached the same problem

Problem Definition: DSN Version



Problem Definition: SC Version

Helping Libraries – SCR

SCR Example – Checkpoint

```
1   do {
2       /** Update the domain, exchanging information with neighbors */
3       UpdateLocals(locals, NBLOCALS, taskid, numtasks);
4       /** Compute the local error */
5       local_error = LocalError(locals, NBLOCALS);
6       /** Compute the global error */
7       MPI_AllReduce(&local_error, &global_error, 1, MPI_DOUBLE,
8                     MPI_MAX, MPI_COMM_WORLD);
9       SCR_Need_checkpoint(&scr_want_to_checkpoint);
10      if( global_error > THRESHOLD && scr_want_to_checkpoint ) {
11          SCR_Start_checkpoint();
12          f = fopen(scr_file_name, "w");
13          if( NULL != f ) {
14              n = fwrite(f, locals, NBLOCALS * sizeof(double));
15              rc = fclose(f);
16          }
17          SCR_Complete_checkpoint(f != NULL &&
18                                  n == NBLOCALS * sizeof(double) &&
19                                  rc == 0);
20      }
21  } while( global_error > THRESHOLD );
```

Rollback-Recovery: DSN vs. SC

DSN

- Started in the 80's, abstract
- Dependability-centric
  - Focus on the algorithms and techniques for reliability
- Rigorous approach
- Typical success story: A prototype implementation with good performance for some applications
- No longer a "hot" topic

SC

- Started in the 00's, concrete
- Application-centric
  - Want to make an application become more reliable
- Adhoc approach
- Typical success story: A real application that can finish despite failures, working code a must, per-application solution OK
- Very active "research" topic

Resilience Research in SC

Rediscovering and reinventing results of DSN:
Examples:
- PPoPP'17
- ABFT

Recognition of the Problem: Standardization Efforts in MPI 3.0

- To my knowledge, not a single DSN member was involved
- *it is the job of the implementor of the MPI subsystem to insulate the user from this unreliability, or to reflect unrecoverable errors as failures. Whenever possible, such failures will be reflected as errors in the relevant communication call. Similarly, MPI itself provides no mechanisms for handling processor failures.*
- *This document does not specify the state of a computation after an erroneous MPI call has occurred.*

Resilience in MPI 3.0

- Confined to returning error code if the running time faces a problem!
- Focus mostly on communications error, which are quite rare in supercomputing applications
- Does not address how to handle the state of a program, relying on the programmer to do all the hard work
  - Called ULFM: User-Level Failure Mitigation
  - Defined some error codes that are more descriptive

Application Programmer to deal with it

```
#include "mpi.h"
#define IC_CREATE_TAG 100
int main( int argc, char *argv[] )
{
        int i, myrank, size;
        MPI_Comm manager_comm;
        MPI_Init( &argc, &argv );
        MPI_Comm_rank( MPI_COMM_WORLD, &myrank );
        MPI_Comm_size( MPI_COMM_WORLD, &size );
        MPI_Intercomm_create( MPI_COMM_SELF, 0, MPI_COMM_WORLD, 0, IC_CREATE_TAG,
&manager_comm );
        MPI_Comm_set_errhandler( manager_comm, MPI_ERRORS_RETURN );
        MPI_Comm_free( &manager_comm );
        MPI_Finalize( ); exit( 0 );
    }
```

Currently Hot

- ABFT being rediscovered
- Books, papers, tutorials, etc.
- More standardization effort

But in Reality

- DSN declared victory when the concepts were discovered, documented, and evaluated
  - We did not go down the engineering path—not interesting, not conducive to research
- SC is facing the problem and it is failing
  - Approaches not rooted in science, more adhoc and engineering
  - The problem persists and it is getting worse

Conclusions and Issues

- One problem was approached differently by two communities, with neither able to provide a working solution to users of high-performance computing systems
- Is it a unique case?
- Why the communities do not talk to one another?
  - Including SC crowd in DSN via workshops worked for 3 years, but then they appeared to have disappeared..
- Do we pay engineering the respect it is due in DSN?
- What can we do differently?