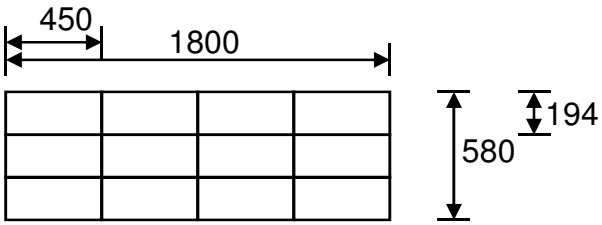
Topology aware Cartesian grid mapping with MPI

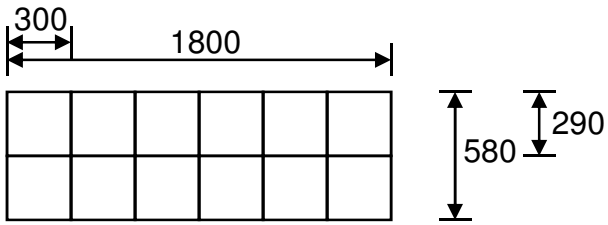
Christoph Niethammer
Rolf Rabenseifner
HLRS

Examples

- Application topology awareness
  - 2-D example with 12 MPI processes and gridsize 1800x580
    - **MPI_Dims_create → 4x3** ☹
    - **grid aware → 6x2** ☺ **processes**



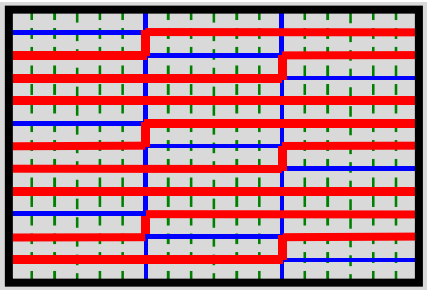
Boundary of a subdomain = 2(450+194) = **1288** ☹



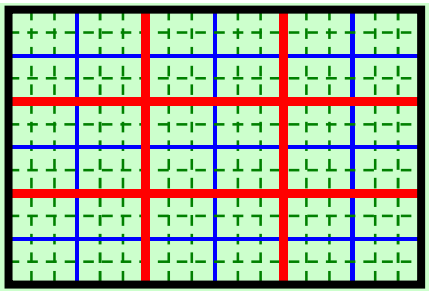
Boundary of a subdomain = 2(300+290) = **1180** ☺

- Hardware topology awareness
  - 2-D example with 9 nodes x 4 CPUs x 6 cores → 12*18 Cartesian processes

Without re-numbering:
150 node-to-node ☹
72 CPU-to-CPU
180 core-to-core



With new ranks:
60 node-to-node ☺
90 CPU-to-CPU
252 core-to-core

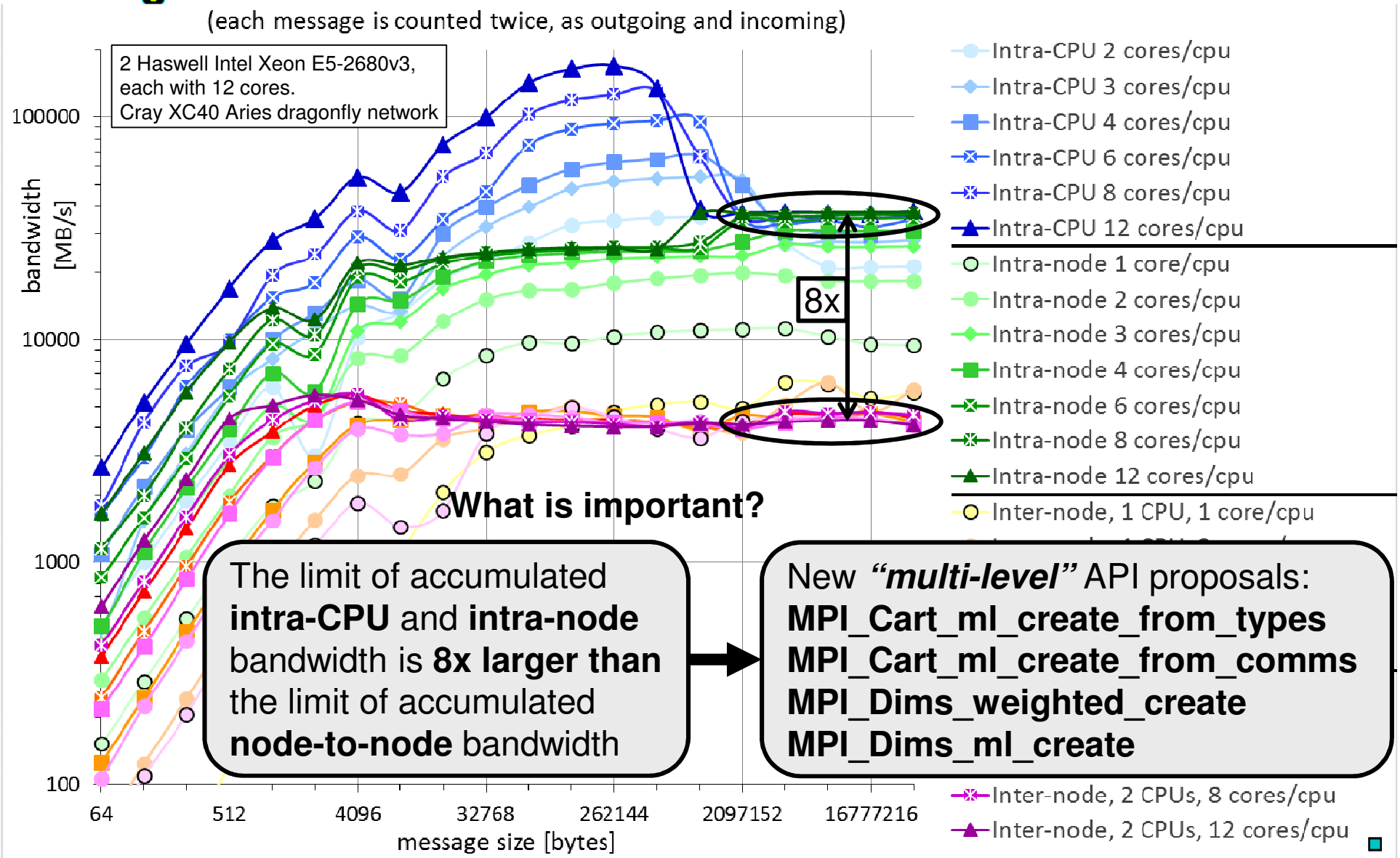


- MPI_Dims_create is limited, e.g., 3-D on 625 nodes x 2 CPUs x 12 cores
  → 25 x 25 x 24 ☹ processes
  → We do
    - Node level: 625= 5 x 25 x 5
    - CPU level: 2= 2 x 1 x 1
    - Core level: 12= 3 x 1 x 4

Result (product): **30 x 25 x 20** ☺

Slides for the poster at EuroMPI 2018
Further info, see https://fs.hlrs.de/projects/par/mpi/EuroMPI2018-Cartesian/

Duplex accumulated ring bandwidth per node



The limit of accumulated **intra-CPU** and **intra-node** bandwidth is **8x larger than** the limit of accumulated **node-to-node** bandwidth

→

New ***"multi-level"*** API proposals:
MPI_Cart_ml_create_from_types
MPI_Cart_ml_create_from_comms
MPI_Dims_weighted_create
MPI_Dims_ml_create

Measurement with halo_irecv_send_multiplelinks_toggle.c on 4 nodes of Cray XC40 hazelhen.hww.de, June 15, 2018, HLRS, by Rolf Rabenseifner (protocol 10)