

# MPI Forum July 2009

---

Ticket #33 – Fix Graph Topology Interface

... the 3<sup>rd</sup> ☺

Torsten Hoefler



# Problem Statement

- Highly-discussed topic, elevator pitch:
  - Current interface specifies **full** graph at **each** process
  - Process 0 knows all neighbors of process 5
    - Which algorithm requires this?
    - Could be handled much better (scalable) on top of MPI
  - → this is non-scalable!
  - Some bounds:
    - We assume  $P$  processes
    - $O(P^2)$  memory per process –  $O(P^3)$  total
  - Practical considerations:
    - 100 processes, 4 byte per edge:  $S \leq 40$  kiB
    - 1000 processes:  $S \leq 4$  MiB
    - 10000 processes:  $S \leq 400$  MiB



# Our Solution

---

- Add new type `MPI_DIST_GRAPH`
  - Doesn't interfere with `MPI_GRAPH` (MPI-2.2!)
  - $O(P)$  memory per process!
- Two new functions:
  - `MPI_Dist_graph_create_adjacent()`
    - User specifies all in and out-edges at each process
    - $O(1)$  creation overhead
  - `MPI_Dist_graph_create()`
    - User specifies arbitrary edges at arbitrary processes
    - Requires half the memory of `_adjacent()`
    - $O(\log(P))$  creation overhead



# Criticism

---

- Process reordering is hard
  - Yes, it is hard! But it's not harder than with the current interface (can be emulated trivially!).
  - Literature exists, subject to ongoing research.
- “It doesn't help”
  - Yes, it does!
  - $O(P^2)$  vs.  $O(P)$  (the expected case scales similarly)
- “Implementation is unclear”
  - See next slide!



# Implementation Outline

---

- Build vectors of edges for each peer
- Compute number of edges for each peer
- Exchange edge counts (MPI\_Reduce\_scatter(\_block) ☺)
  - Each process knows that it'll receive X edges
- Post X nonblocking receives (ANY\_SOURCE)
- Send all edges
  - Done!  $\Omega(\log(P))$ ,  $O(P^2)$



# More Differences/Features

---

- Edges have weights!
  - we have MPI\_UNWEIGHTED if the user doesn't want weights!
- Creation calls accept an Info object!
  - It's use is not defined → a possibility for vendors to add and test their own metrics!
- Fully downwards compatible!
  - The generalized interface (not adjacent) allows a specification in MPI-2.1 style
  - It's not recommended though!



# Discussion!

---

- I contacted everybody who didn't vote yes
  - I got one reply
  
- Are there any questions/discussions left?
  - We should be absolutely sure before voting!
  - We can go through the ticket again (?)
  
- Special thanks to:
  - Jesper L. Traeff, Rolf Rabenseifner, Bronis de Supinski, and Rajeev Thakur



---

# Voting 😊

---

Let's vote!

