MPWide: A communication library for wide area message passing

Derek Groen
Centre for Computational Science

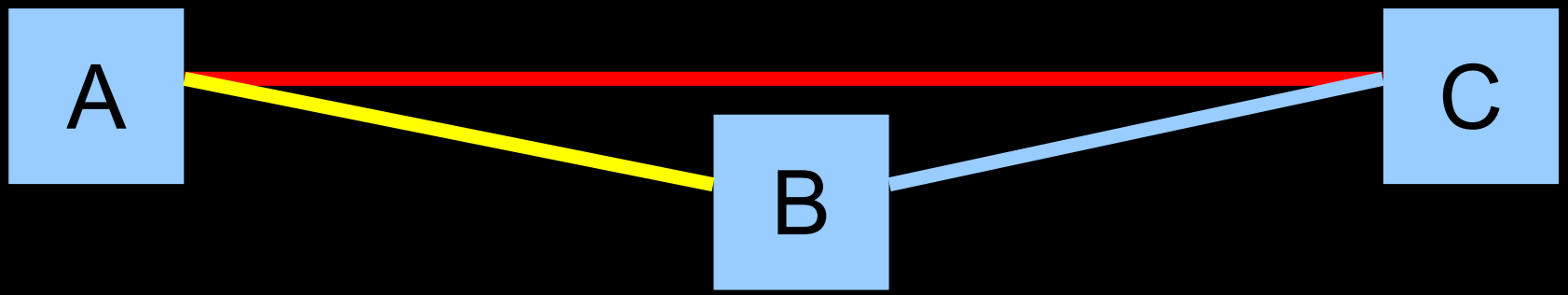
UCL

Overview

- The networking landscape
- Using wide area networks
- MPWide
- Example applications
- Uses for multiscale modelling
- Questions

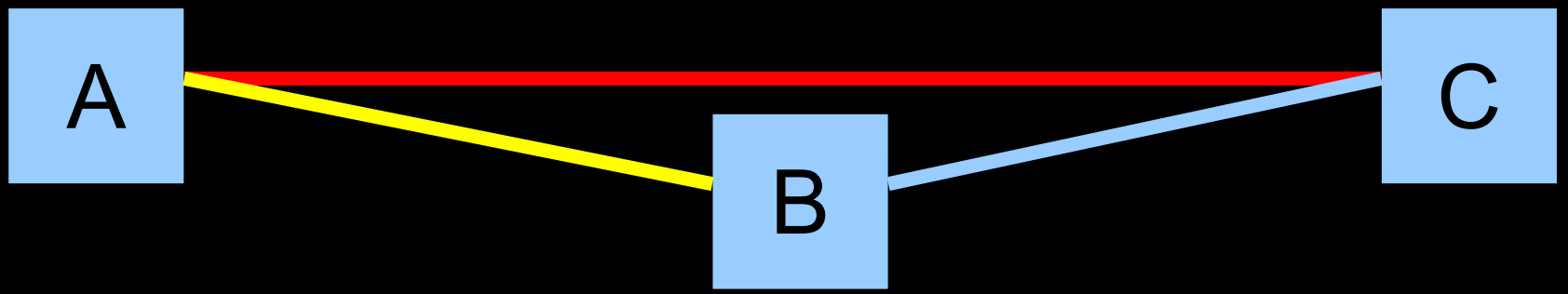
The networking landscape

- The networks connecting grid sites and supercomputers are highly heterogeneous.
  - Configurations differ at end points.
  - Shared paths vs. Dedicated paths
  - Optical interconnects vs. Regular interconnects.



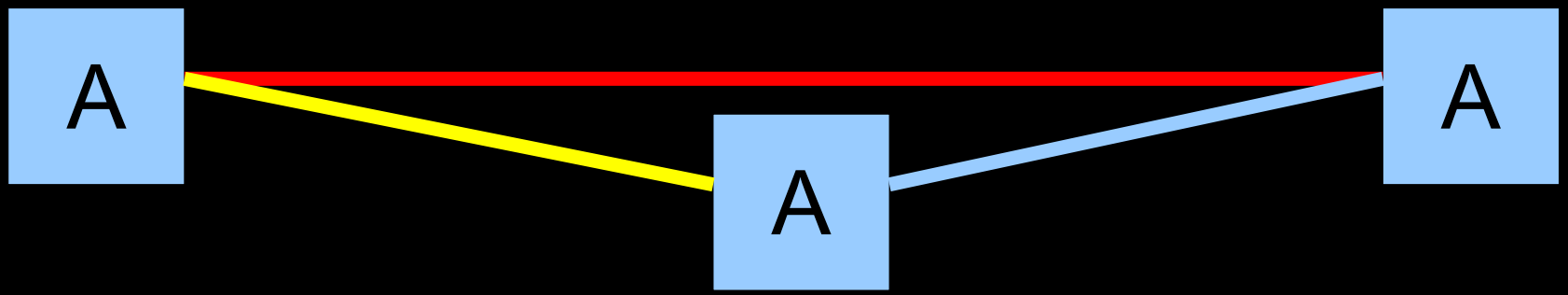
The networking landscape

- Fundamental issue: Networks configurations tend to be *node*-specific, not *path*-specific.
  - What do we do when a node has multiple paths?
    - (most nodes nowadays do)



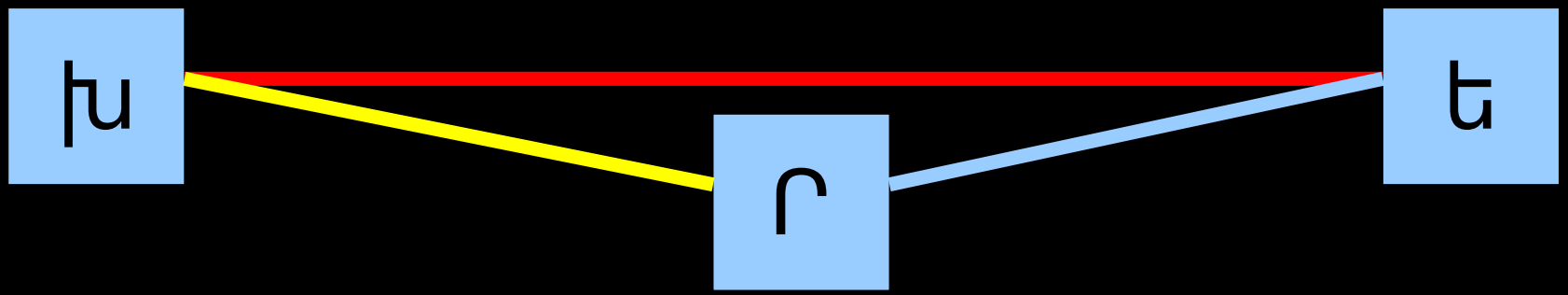
Using wide area networks (WANs)

- Solution 1: Apply a homogeneous configuration for all paths.
  - Could work for nodes with similar path lengths.
    - Not common for WAN communication nodes.
  - Inefficient for the TCP protocol, where the optimal config is dependent on the path length.
  - Requires admin privileges on all end-points.



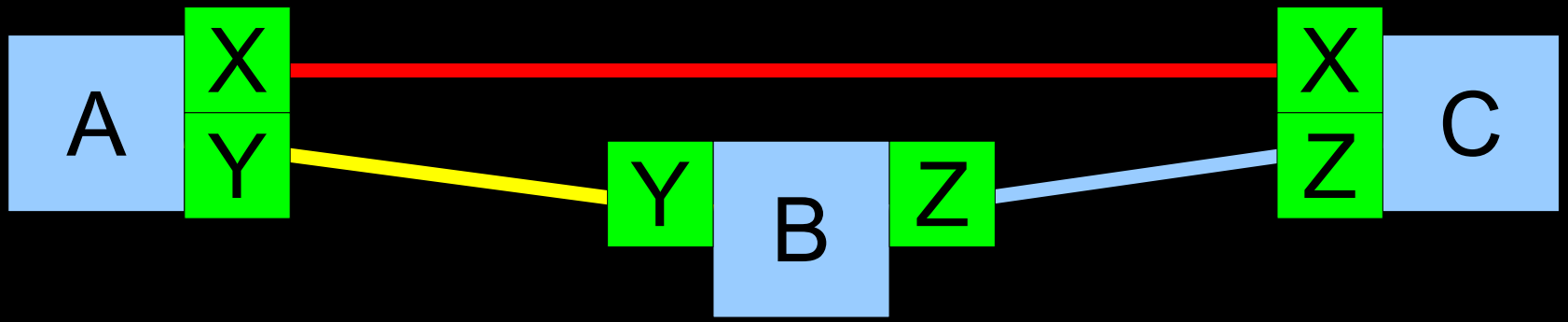
Using WANs

- Solution 2: Adopt a different protocol.
  - May accomodate heterogeneous configs.
  - New protocol, new list of potential issues.
  - Interplay between protocols on shared networks.
  - Time-consuming and politically heavyweight process.



Using WANs

- Solution 3: User-space tuning through software.
  - Limited space for tuning.
    - Some adjustments require admin rights.
  - Use TCP protocol and existing configurations.
  - No special privileges required.

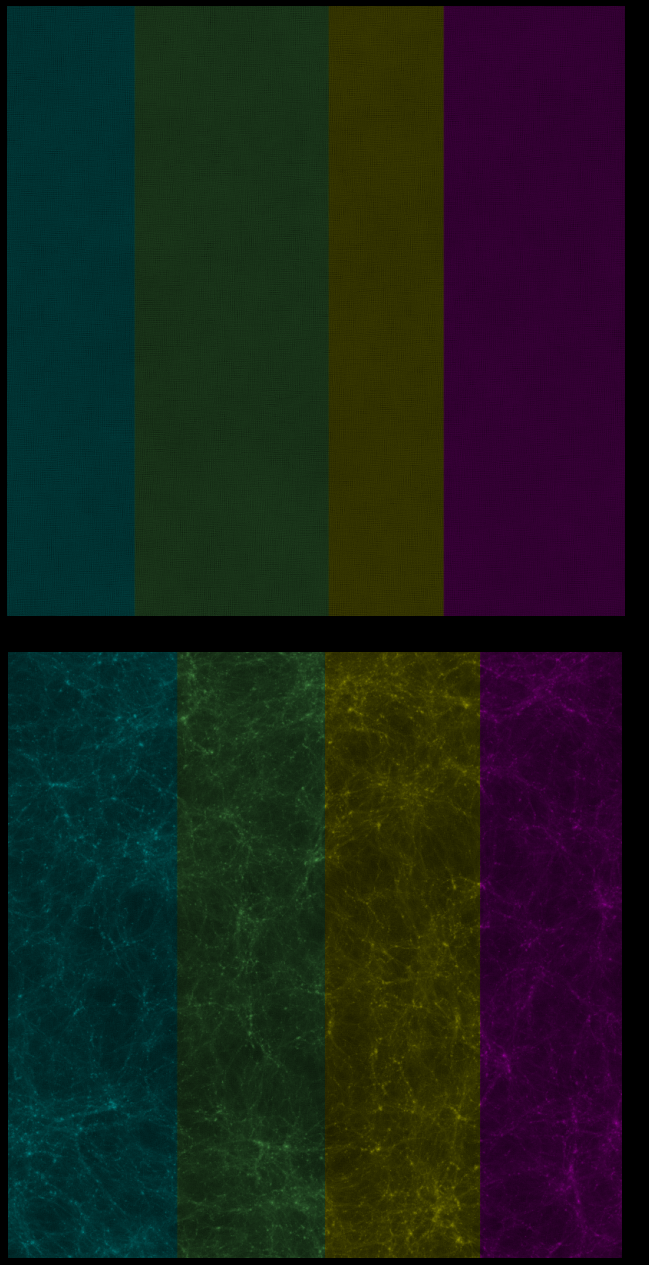


MPWide

- MPWide is a communication library which allows for user-space tuning of individual paths.
- For each path it can:
  - Use 1 or multiple tcp streams.
    - Good performance obtained with up to 128 streams/path.
  - Configure different buffer and packet sizes.
  - Apply software-based packet pacing to reduce load.
    - Also improves performance on long networks (Yoshino et al. 2008).

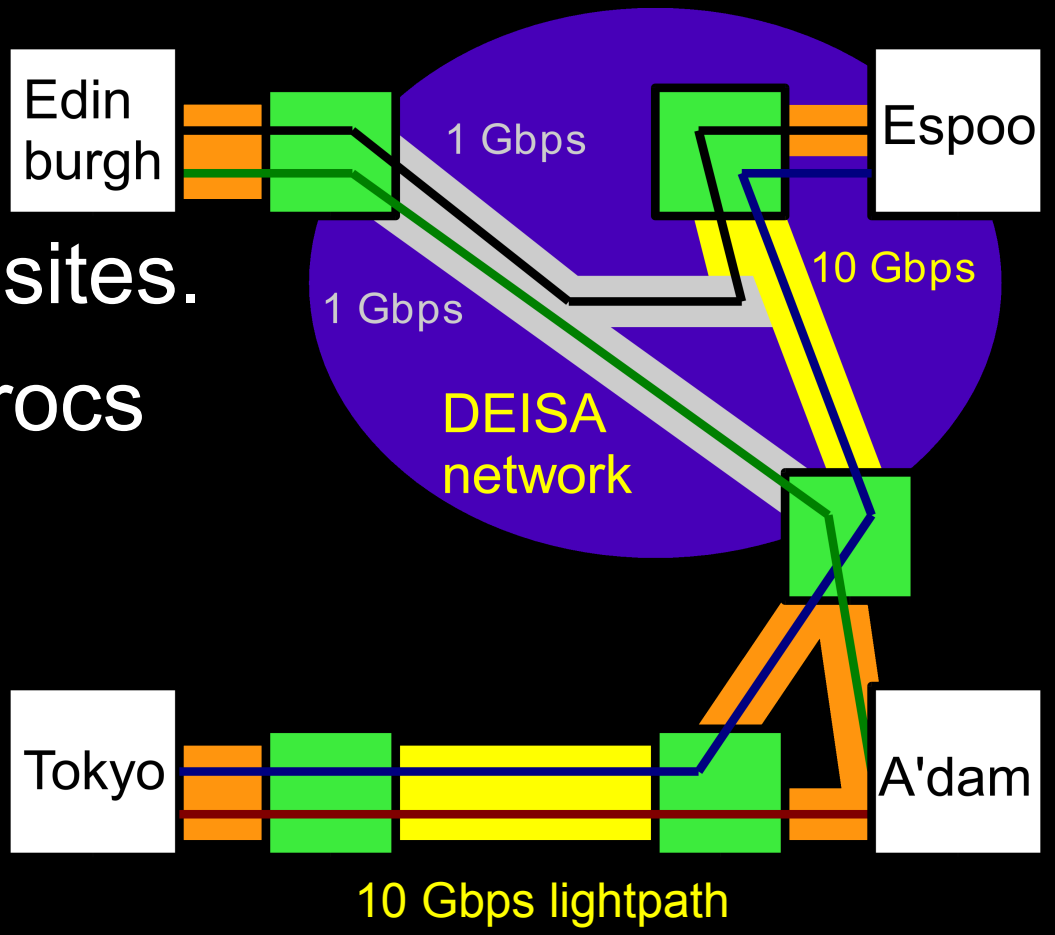
Example: cosmological N-body

- One simulation, parallelized across supercomputers.
- Uses the SUSHI code, which is a cross-site adaptation of GreeM.
- Models dark matter structure formation over 13.4 billion years.
- Algorithm: Tree + Particle-mesh.
- Adaptive load-balancing between sites. →

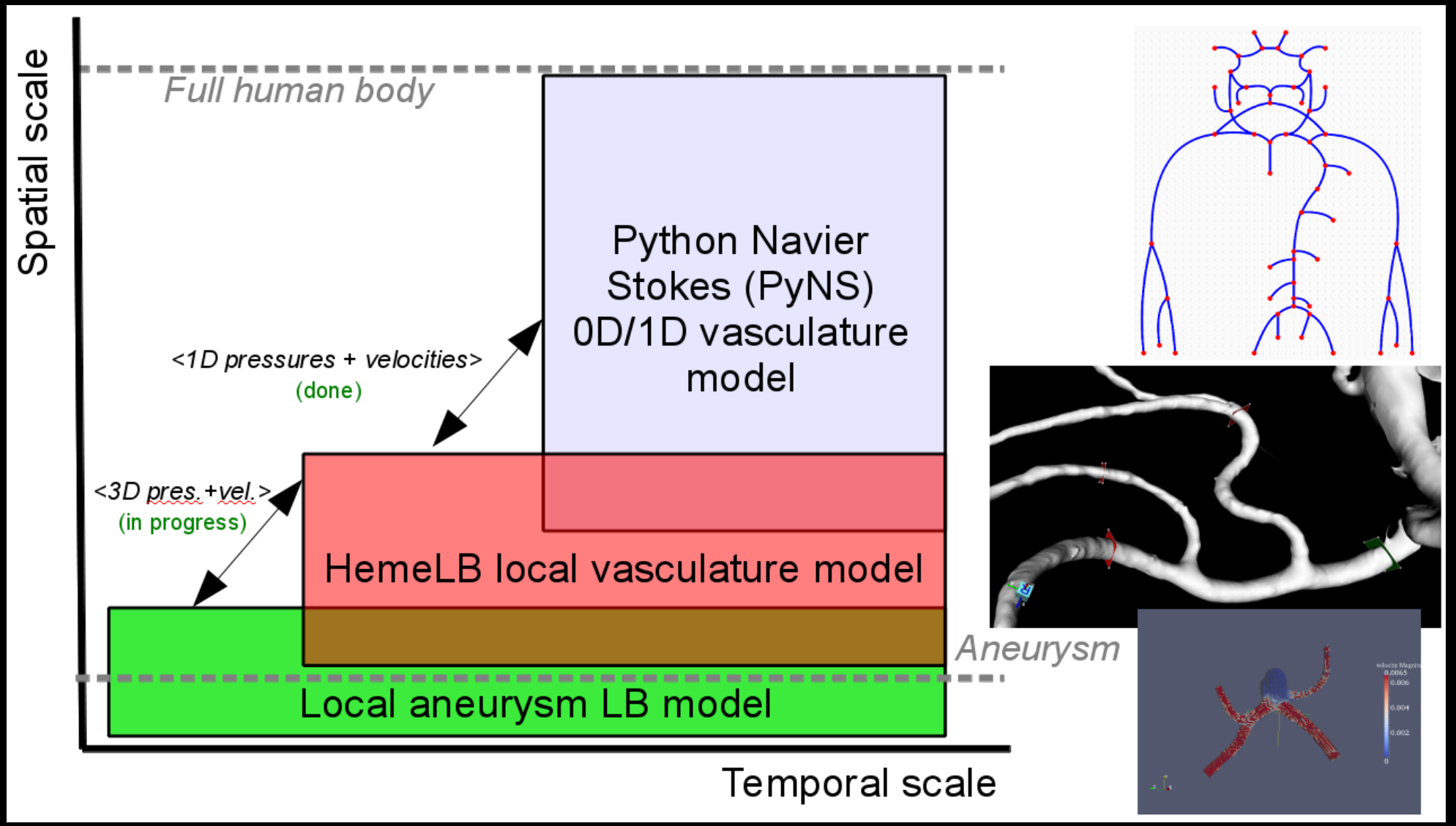


Example: cosmological N-body

- Using 2 to 4 supercomputers simultaneously.
  - Up to 2048 cores total.
- MPI within each site.
- Custom MPWide connections between sites.
- MPWide *Forwarder* procs bypass connectivity restrictions.
- 2048 cores, 3 sites, 7% comm. overhead.

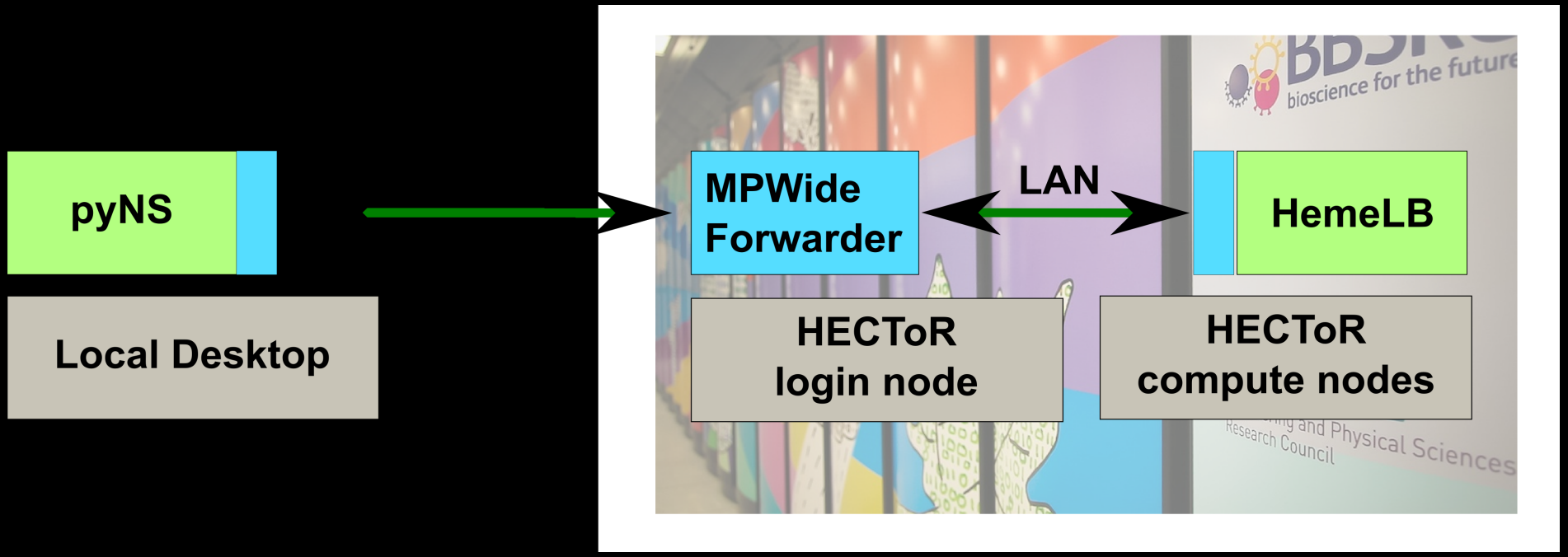


Example: multiscale bloodflow



Example: multiscale bloodflow

- pyNS (1D) coupled to HemeLB (3D).
- 400.000 time steps, 4000 velocity exchanges
  - with 1.2% comm. overhead (512+1 cores, 2298 s),
  - and 5% comm. overhead (2048+1 cores, 907 s.).



Uses for multiscale modelling

- Can be used for performance critical cyclic coupling over wide area networks.
  - High-performance, simple low-level interface.
- Contains an *mpw-cp* file transfer client to accelerate file-based couplings.
- Supports C, C++, Python.
- Trivial to install and intended for users without administrative privileges.
- Is being integrated into MUSCLE 2 to improve its coupling performance.

Thank you!

- MPWide website:
  - http://castle.strw.leidenuniv.nl/software/mpwide.html
- More on the multiscale bloodflow application:
  - Groen et al., Interface Focus 3(2), 2013.
- More on the cosmological N-body application:
  - Groen et al., INFOCOMP 2011, ArXiv:1109.5559.
- Thanks go out to Steven Rieder, Simon Portegies Zwart, Tomoaki Ishiyama, Keigo Nitadori, Joris Borgdorff, Rupert Nash and the MAPPER consortium as a whole.