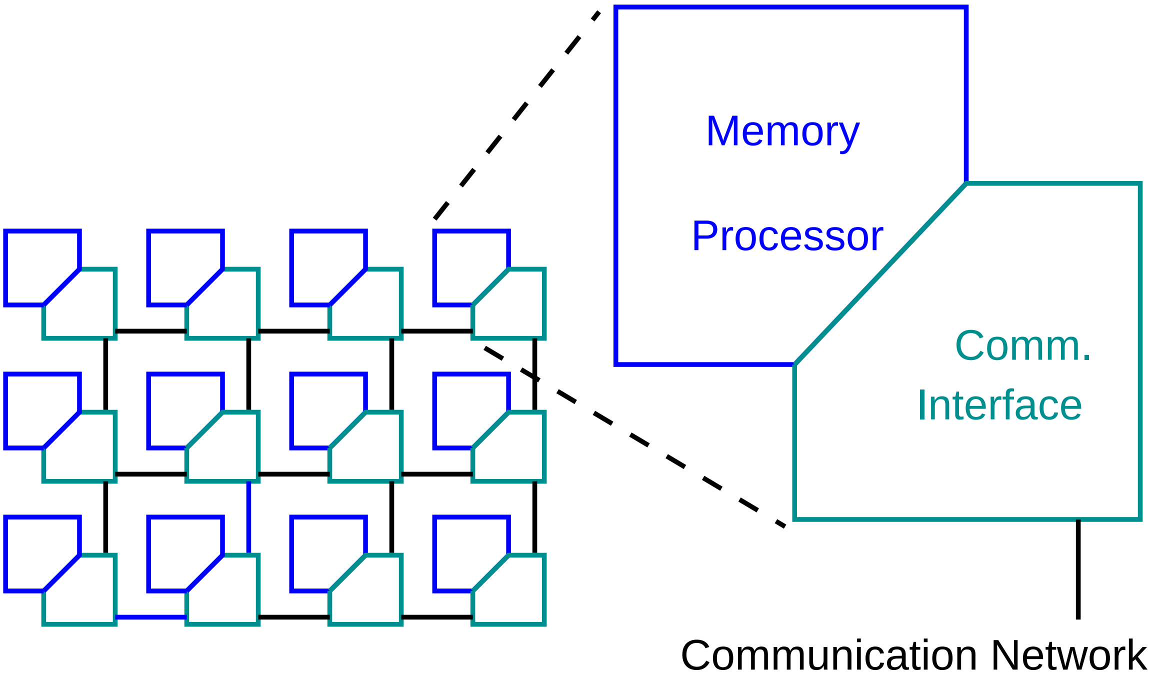
Message Passing Programming I

Wolfgang Schreiner
Research Institute for Symbolic Computation (RISC-Linz)
Johannes Kepler University, A-4040 Linz, Austria

Wolfgang.Schreiner@risc.uni-linz.ac.at
http://www.risc.uni-linz.ac.at/people/schreine

Distributed Memory Architectures

- Multicomputers,
- Loosely coupled architectures,
- Message passing architectures,
- MPP Massively parallel processors.



- Local memories private to processors,
- Communication via network.

System Interconnect Architectures

Characteristics

- Degree

  Maximum number of links per node.

- Diameter

  Maximum distance between two nodes

- Bisection width

  Minimum number of edges to be removed to divide network into halves.

- Edge length

  Physical representation in 2D/3D.

Network topology.

System Interconnect Architectures

Performance

- Functionality

  Data routing, global synchronization, ...

- Latency

  Maximum time delay for message transfer.

- Bandwidth

  Maximum data transfer rate.

- Hardware complexity.

  Implementation costs.

- Scalability

  Expandable with more nodes.

Network technology.

Static Connection Networks

See additional Figures.

- k-ary n-cube networks: dimension k, $N = n^k$ nodes.
  - Line/ring (1D), mesh/torus (2D), cube (3D).
  - Degree: 2^k = const.
  - Diameter: $k(n-1) = k\sqrt[k]{N}$.
  - Bisection width: n^{k-1}.
  - Edge length: constant.
- Hypercube: binary n-cube, $N = 2^n$.
  - Degree: n *(not constant).*
  - Diameter: $n = \log_2 N$.
  - Bisection width: $2k^{k-1}$.
  - Edge length: not constant.
  - Scalability only by factors of 2!

Today, low-dimensional (2D, 3D) meshes are preferred over hypercubes.

Static Connection Networks

- Trees: degree k, depth d, $N = k^d - 1$ nodes.
  - Degree: $K + 1$.
  - Diameter: $2(d - 1) \approx 2 \log_k N$
  - Bisection width: 1! *(hot spot!)*
  - Edge length: not constant.
  - Root becomes communication bottleneck.
- Fat Tree
  - "Branches get thicker toward the root".
  - Channel width increases toward the root.
  - Communication bottleneck is avoided.
  - Example: Connection Machine CM-5.

Fat trees are also popular topologies.

Dynamic Connection Networks

Switches provide dynamic connectivity.

- Digital buses.

  Only one transaction at a time.

- Crossbar network.

  All-to-all connection by n^2 switches.

- Multistage networks.

  Multible levels of $a \times b$ switches; several kinds of inter-stage connection patterns possible.

- Multi-dimensional networks.

  Multiple dimensions of buses or crossbar switches.

Communication patterns implemented on demand.

Message Formats

Message = variable-sized block of data = $\langle$destination, address, size$\rangle$.

- Messages = application-level concept.

  Computation processor inititates communication by passing necessary information to communication processor.

- Packages = system-level concept.

  Communication processor splits message into fixed size packages and transfers them over the network.

- Flits (flow control digits).

  Packages are split into fixed-length flits that may flow through the network in pipelined fashion.

Message Routing Schemes

- Store-and-Forward Routing
  - Packages are units of information flow.
  - Each package is stored in intermediate node before it is forwarded to the next node.
  - Latency proportional to communication distance.
- Wormhole Routing
  - Flits are units of information flow.
  - Each received flit of a package is forwarded to next node before the next flit is received.
  - Latency almost independent of communication distance.

Modern multicomputers use wormhole routing!

Wormhole Routing

Transmit a 100 byte message through a 100 MB/s network.

- 1 μs transmission time.
- Store and forward:
  - 1 nodes: 1 μs.
  - 8 nodes: 8 μs.
- Wormhole: 4 bytes/flit.
  - 1 node: 1 μs.
  - 8 node: $1 + \frac{7}{25} = 1.28$ μs.

Distance of communication does not matter any more!

Routing Strategies Characterization

- Deterministic routing.

  Communication path solely determined by source and destination address.

- Adaptive routing.

  Communication path depends on network conditions; alternate paths are possible.

- Non-deterministic routing.

  Random path is selected to avoid hot spots.

Minimization of communication distance is less important than avoiding hot spots in the network!

Dimension-order Routing

- Deterministic strategy for multidimensional networks.
- Choose specific dimension order and route along pathes in that order.
- Meshes: X-Y routing.

  First route along X-dimension, then route along Y-dimension.
- Hypercubes: E routing.

  Route along dimensions in the order of different bit positions of source and destination address.

Dead-lock free routing schemes.

Broadcasting and Multicasting

Communication Patterns

- Broadcast pattern

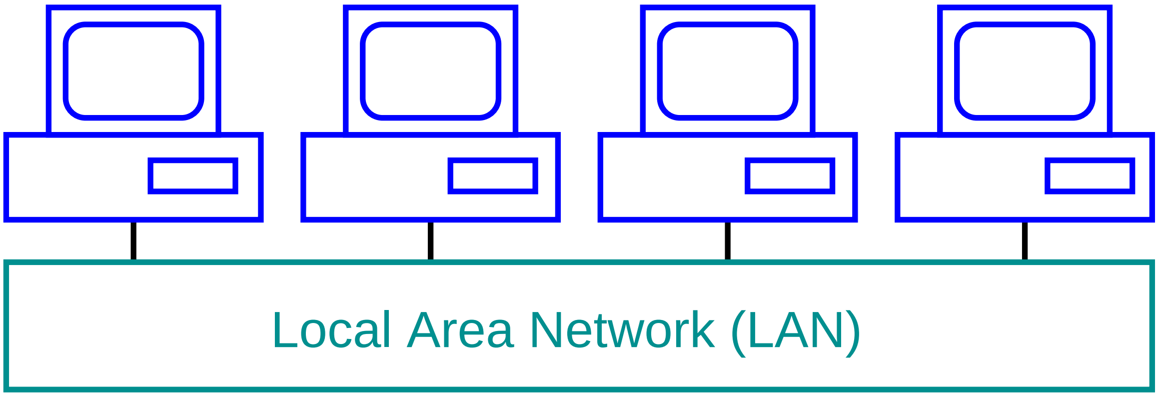
  One-to-all communication; one source sends the same message to all destinations.

- Multicast pattern

  One-to-many communication; one source sends same message to multiple destinations.

Avoid multiple unicasts by utilizing multicast routing schemes!

Computer Clusters



Multiple computers connected via network.

- Nodes are conventional workstations.
- Network is conventional LAN.
- Advantage: low costs.
- Disadvantage: high latencies.
- PVM: Parallel Virtual Machine

  Message passing package for clusters.

Basic operation like MPP; fault-tolerance must be considered.