



An Invited Panel Presentation to Workshop on HPC Programming Languages:

# What is a Realistic Vision of the Future

Thomas Sterling

Department of Computer Science

December 13, 2006



AT LOUISIANA STATE UNIVERSITY



**cct**

email: "recent board picture"





email: "recent board picture"



The whiteboard contains several diagrams and notes:

- Top Left:** Two 4x4 grids of squares, one above the other.
- Top Center:** A diagram showing a central square with an 'X' inside, surrounded by eight smaller squares. To its left is a diagram of a multi-layered structure with a 'PE' label.
- Top Right:** A list of four numbered points:
  - ① How many mesh pairs?
  - ② Area of DA x wrt whole chip
  - ③ Outside connections
  - ④
 Below this list are two small diagrams: one showing a square with 'PE' and 'DA' labels, and another showing a square with 'DA' and 'PE' labels.
- Middle:** A large blue circle containing three numbered points:
  - ① Communication overhead in
    - external
    - internal
 - gilgamesh2
  - ② Memory Size
  - ③ Peak floating point ops/accelerator
- Bottom Left:** A diagram of a square grid with red lines connecting nodes, representing a mesh or communication network.
- Bottom Center:** A graph with 'Peak FLOPS' on the y-axis and '% Accelerator to area' on the x-axis. A blue curve starts high and decreases as the x-axis value increases. A red curve starts low and increases as the x-axis value increases. A vertical dashed line is labeled 'mem capacity'.
- Bottom Right:** A diagram of a grid of squares with some squares shaded. To its right, text reads:
  - #9 per accelerator
  - Mind element - 16 GHz
  - Accelerator - 32 GHz
 Below this text are two small diagrams: one showing a 2x2 grid of squares and another showing a 2x2 grid of circles.



**cct**

It is a realistic vision of the future:



- Power will dominate HPC
- Architectures defined to be for scalable parallel execution
- Memory capacity, latency, and bandwidth to define system scale
- Thousand lightweight cores per socket
- Multimodal leads to heterogeneity
- Billion-way parallelism
- Hardware overhead bounds program parallelism and system scalability
- Chip to chip optical communication (1 Tbps/channel)
- New execution model will extend beyond (not erase) current methods to enable and exploit future technology



# A Realistic Future Vision



