Multiterabit Switch Fabrics Enabled by Proximity Communication

Hans Eberle, Alex Chow, Bill Coates, Jack Cunningham, Robert Drost, Jo Ebergen, Scott Fairbanks, Jon Gainsley, Nils Gura, Ron Ho, David Hopkins, Ashok Krishnamoorthy, Jon Lexau, Wladek Olesinski, Tarik Ono, Justin Schauer

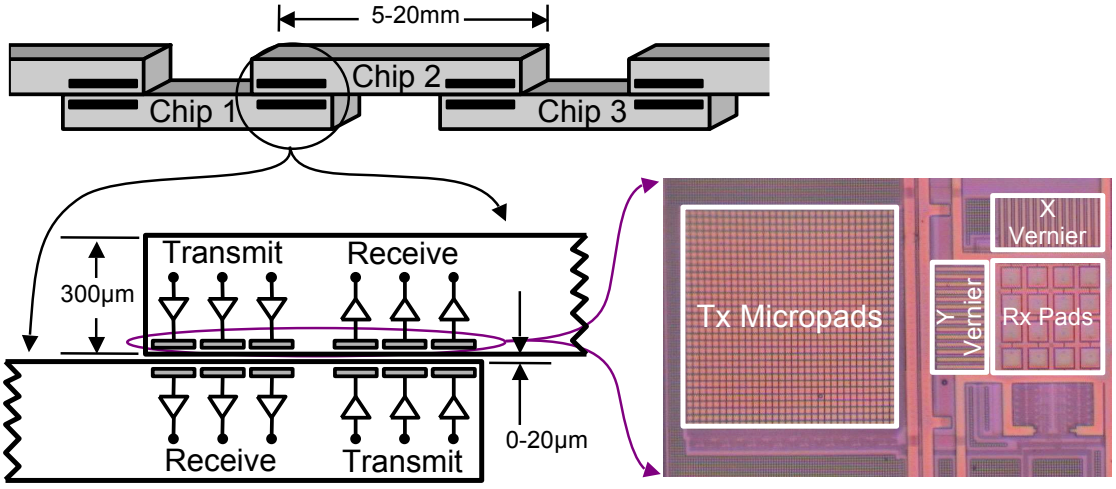
Sun Microsystems Laboratories

Future Interconnect Needs

- The interconnect becomes an increasingly critical system component
  - > Fatter compute nodes
  - > Increasing disparity between local and remote communication
- Data center trends
  - > Server consolidation
  - > Network consolidation
  - > Virtualization
  - > Clustering
  - > Horizontal scale beyond the chassis

Proximity Communication (PxC)

Chips overlap face-to-face

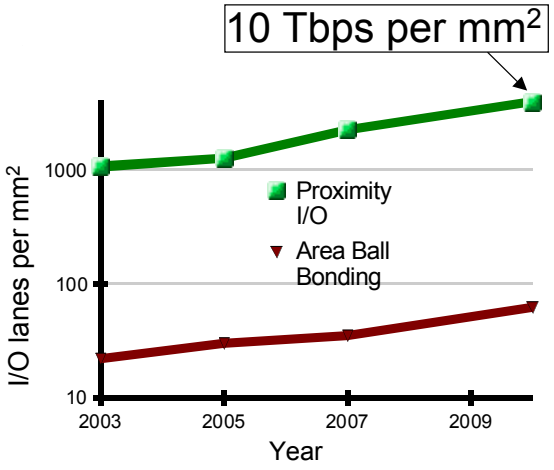


Capacitively couple over micron distances

Utilize on-chip electronic alignment

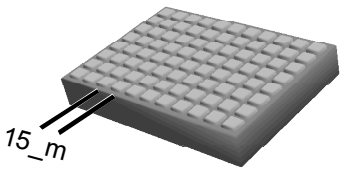
Removing the Chip IO Bottleneck

Huge Bandwidth Gain

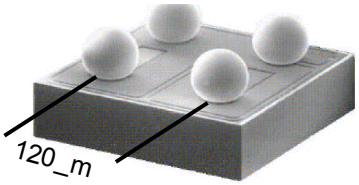


Comparison of Scale

Proximity Communication



Area Ball Bonding

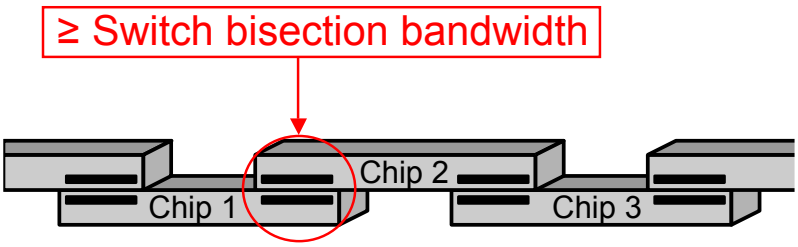


Proximity Communication Advantages

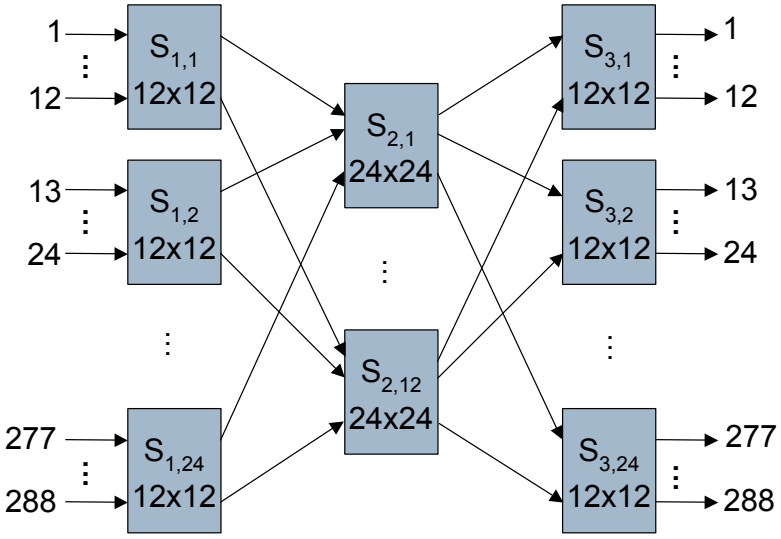
- Increases bandwidth/area
- Avoids off-chip wires
- Obviates ESD protection
- Shrinks transceiver circuits
- Lowers power consumption
- Makes multi-chip modules reworkable
- Enables smaller chips

Opportunity

- Proximity Communication allows for building switch fabrics that scale to thousands of ports and multiple Tbps throughput using a *flat single-stage* network rather than a hierarchical multi-stage network



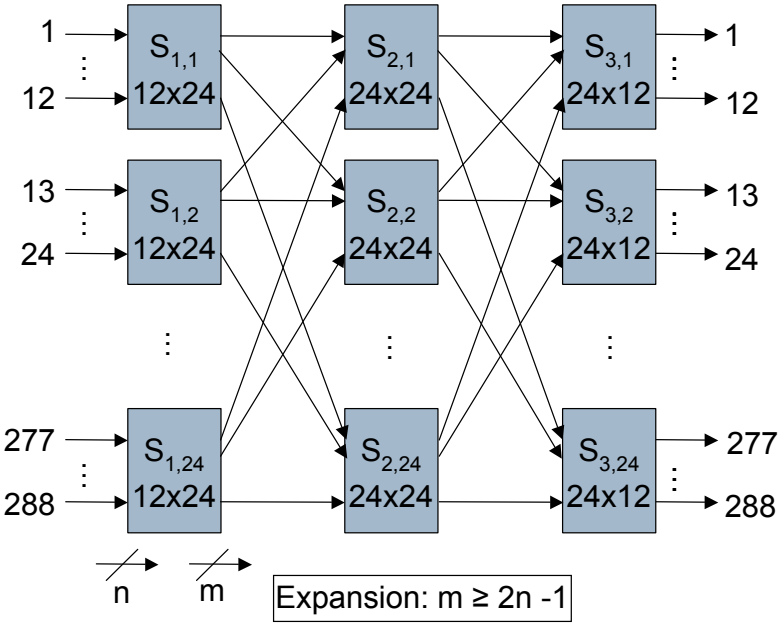
Blocking Multi-stage Switch



(Folded network combines S_1 and S_3)

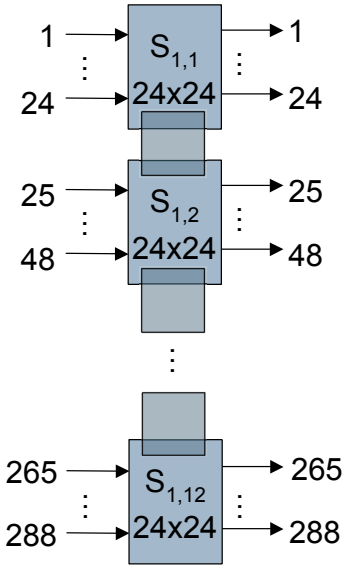
- 36 switches
- 3 stages
- 576 internal links

Non-blocking Multi-stage Switch



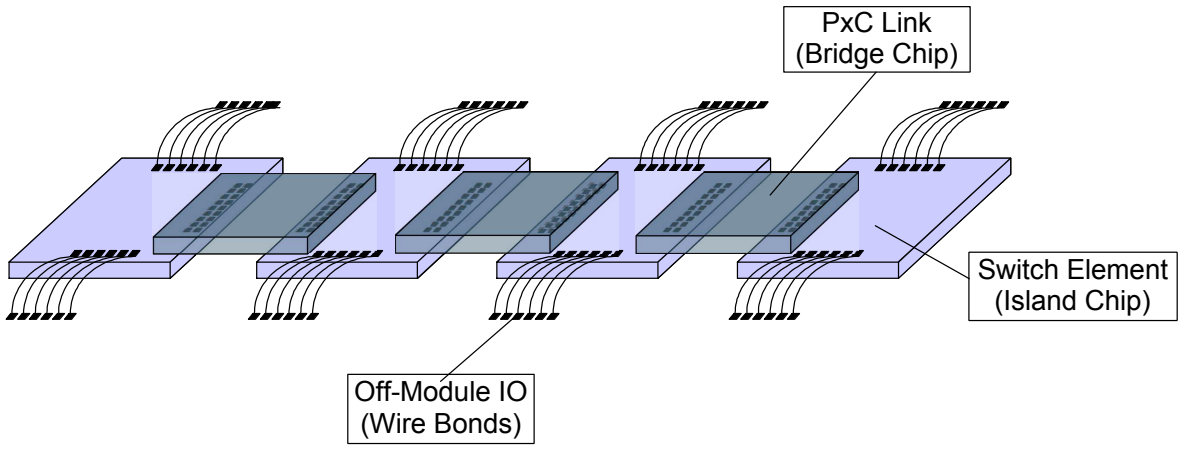
- 72 switches
- 3 stages
- 1,152 internal links

Proximity Communication Switch

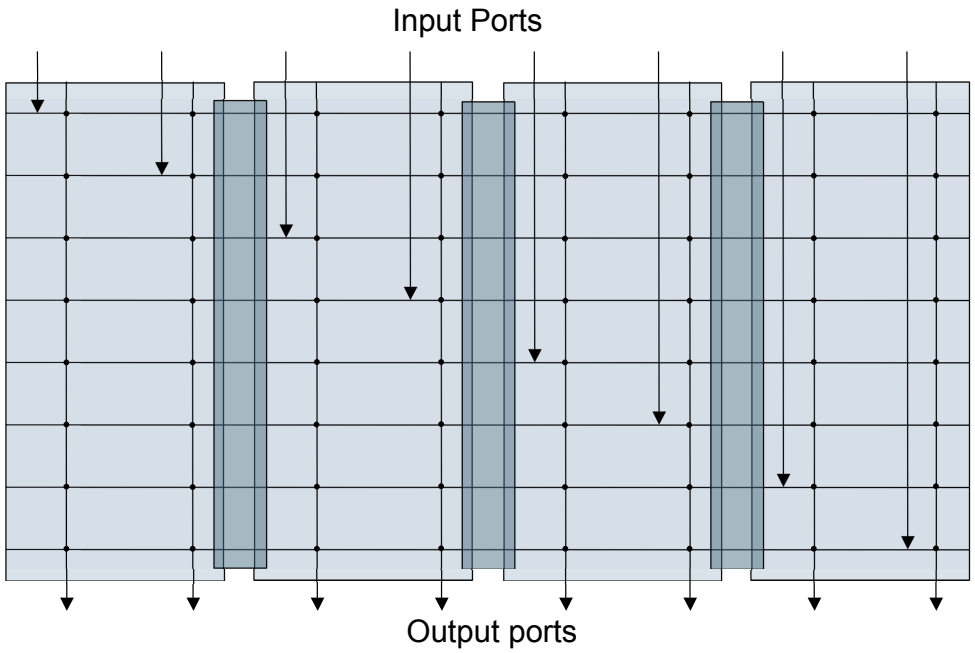


- 12 switches
- 1 stage
- PxC links

Vector Multi-Chip Module



Port-Sliced Crossbar Switch



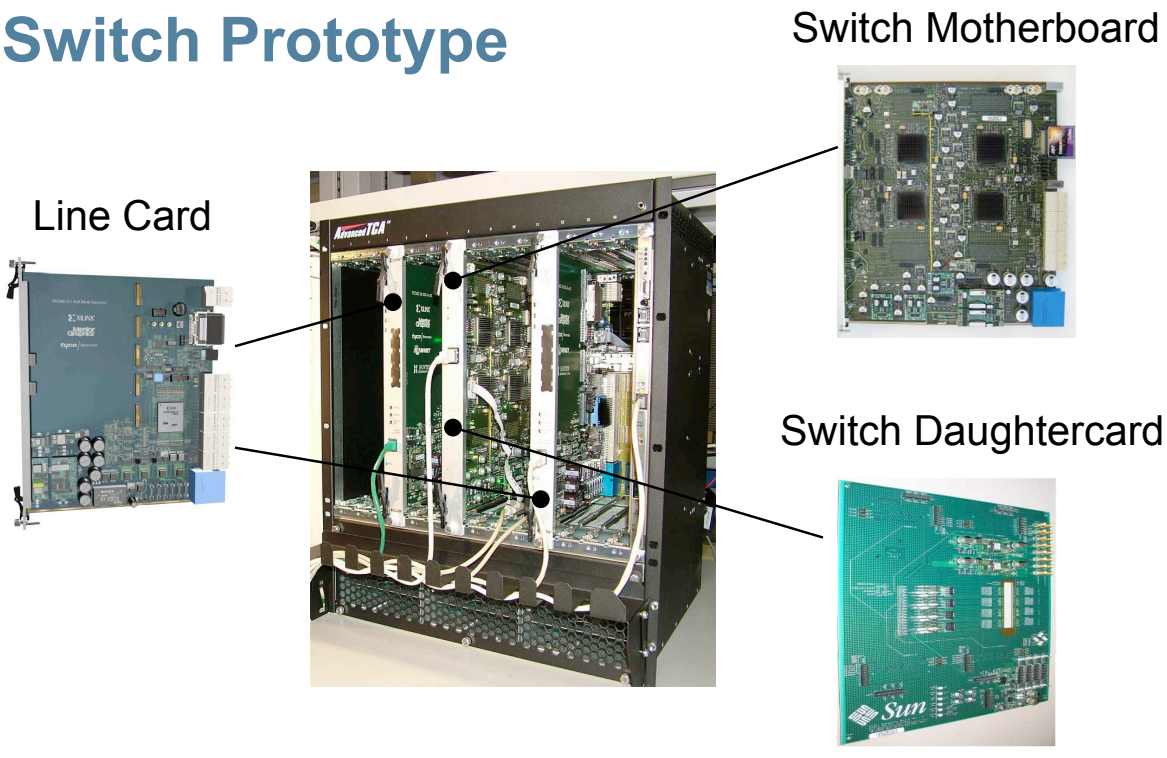
Single-Stage PxC Switch Advantages

- Low deterministic latency
- Simple global scheduling
  - No internal blocking
  - No out-of-sequence delivery
  - Service guarantees possible
- Lower cost
  - Fewer switch elements
  - Less internal wiring
- Less power
- Higher reliability

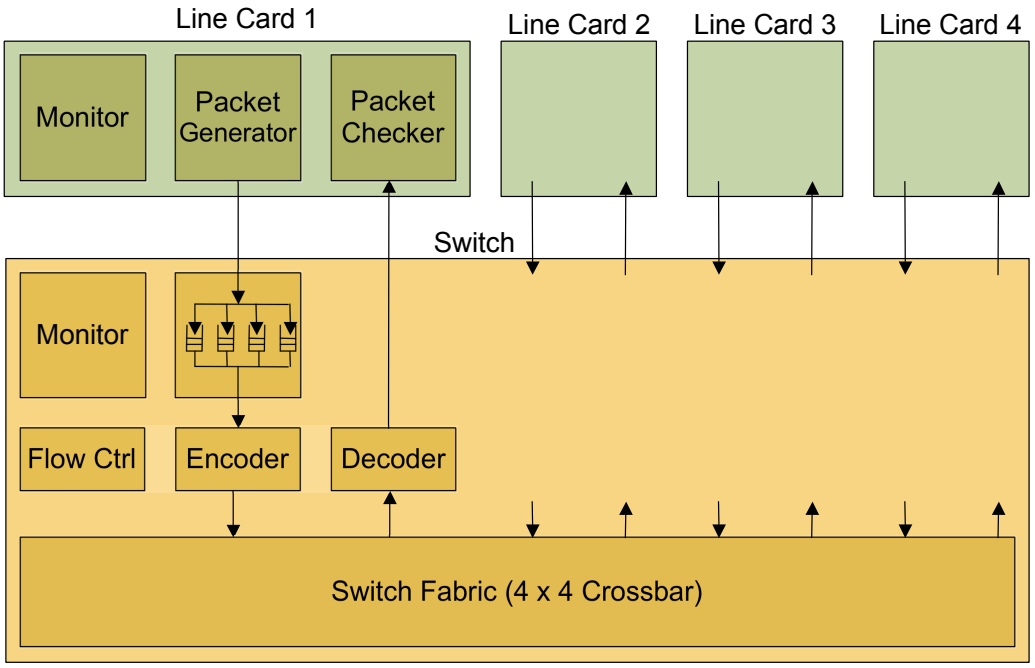
Switch Prototype Characteristics

- System characteristics
  - > 4 x 10GE ports
  - > Layer2 switching
  - > Based on ATCA standard
  - > Off-the-shelf line cards
  - > Proprietary switch blade
- Switch fabric
  - > "Vector switch" with 4 Island chips + 2 Bridge chips (3 PxC links)
  - > Off-chip connections through wire bonds

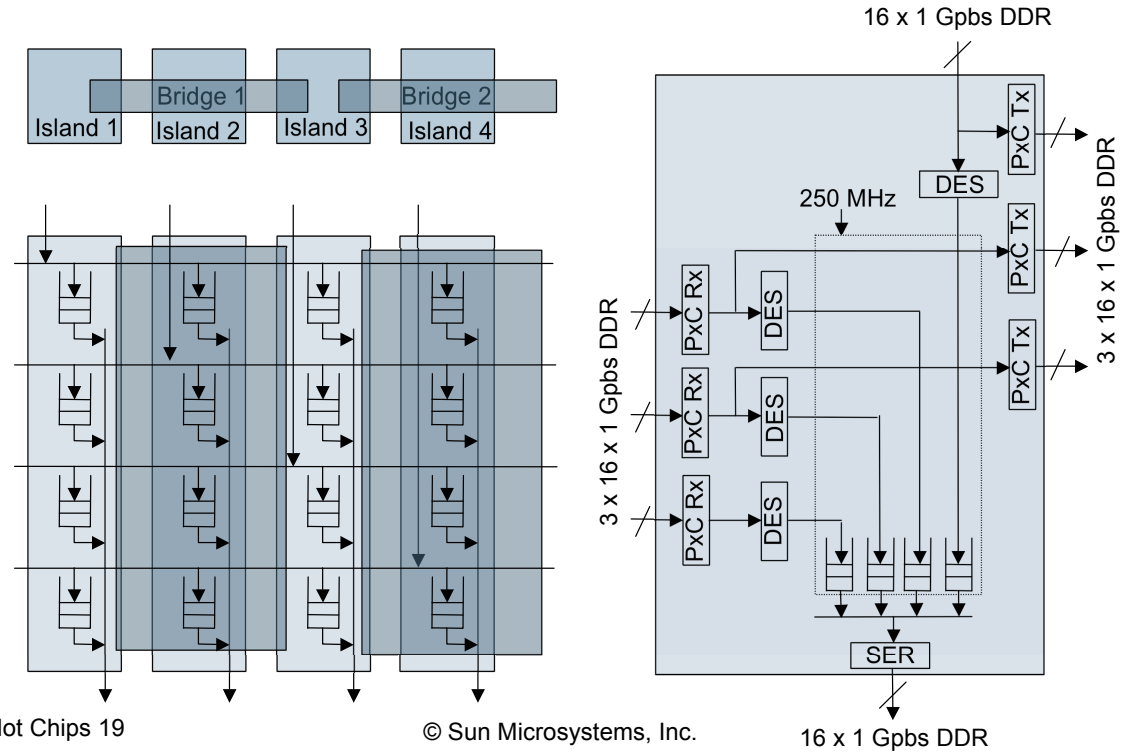
Switch Prototype



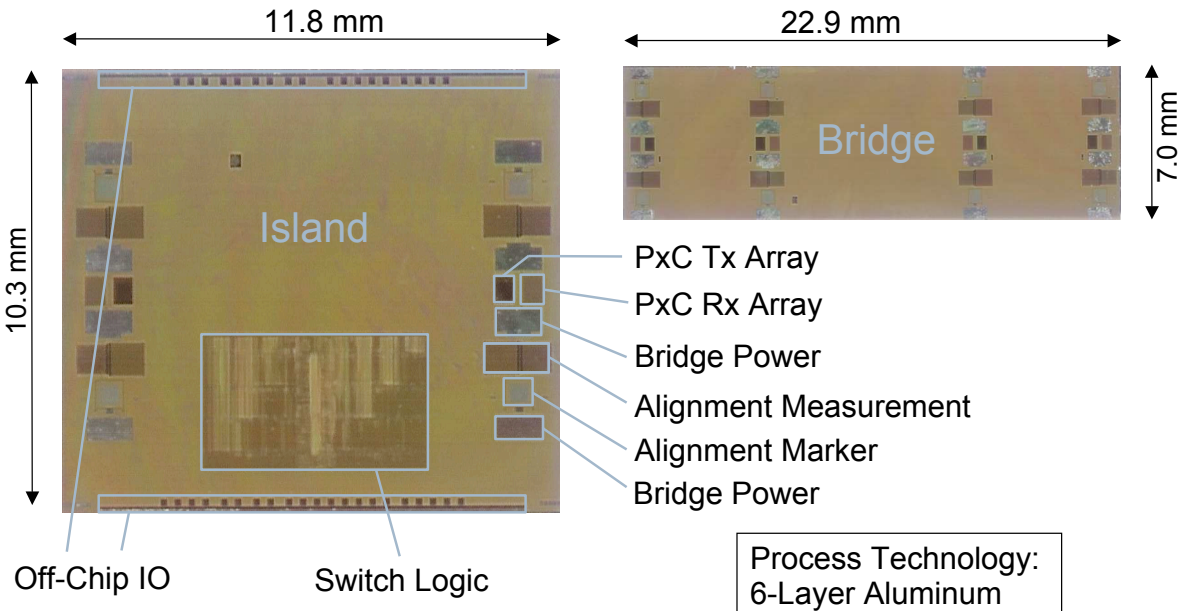
Switch Prototype Organization



Bridge and Island Chips

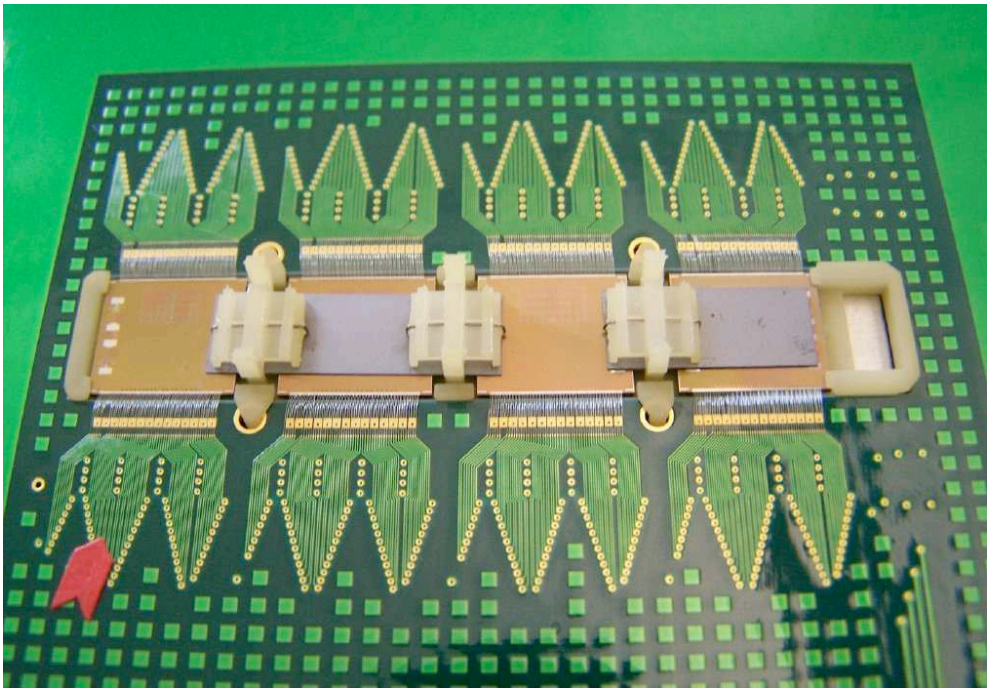


Bridge and Island Chips

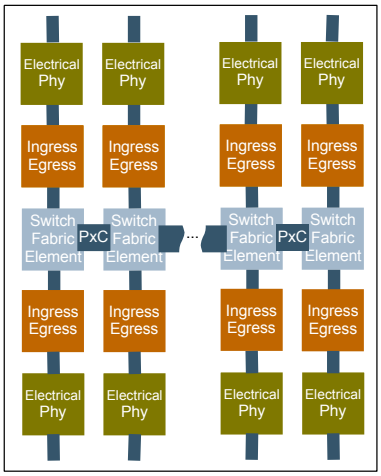


Process Technology:
6-Layer Aluminum
TSMC 0.18µm CMOS

Vector Switch Prototype



Scaling Up



256 Ports
2.5 Tbps

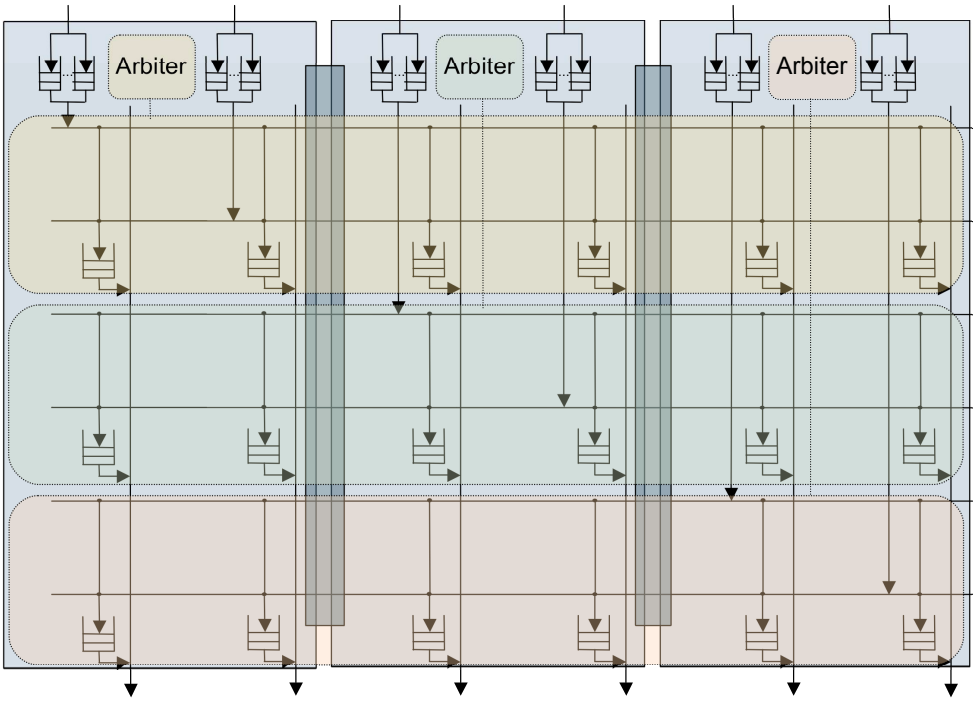
1,024 Ports
10 Tbps

4,096 Ports
40 Tbps

Scalable Switch Architecture

- "Output Buffered Switch with Input Groups“
  - > Reduces memory requirements from $O(n^2)$ to $O(n \bullet \#\ \text{Island Chips})$
  - > To be presented at Globecom 2007
- "Parallel Wrapped Wave Front Arbiter"
  - > Increases throughput of n x n Wrapped Wave Front Arbiter by a factor of n
  - > Presented at HPSR 2007

Output Buffered Switch with Input Groups



Applications

- Data center backbone
- Blade system interconnect
- ATCA chassis aggregation
- Cluster interconnect
- System interconnect

Summary

- Proximity Communication allows for building a *flat single-stage* switch fabric that scales to thousands of ports and multiple Tbps throughput
  - > Low latency
  - > High efficiency
  - > Service guarantees
  - > Low power
  - > High physical density

Hans Eberle
hans.eberle@sun.com