Scalable Interconnects for Reconfigurable Spatial Architectures

Stanford University

Yaqi Zhang yaqiz@stanford.edu
Alexander Rucker acrucker@stanford.edu
Matthew Vilim mvilim@stanford.edu
Raghu Prabhakar raghup17@stanford.edu
William Hwang williamhwang@stanford.edu
Kunle Olukotun kunle@stanford.edu

STANFORD DAWN

Motivation

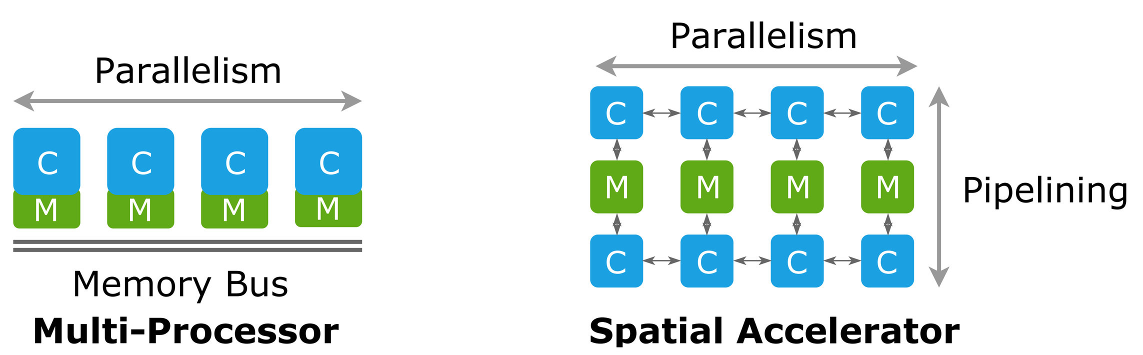
Spatial Accelerators

- Energy efficient
- High-throughput
- Low-latency

Network Requirements

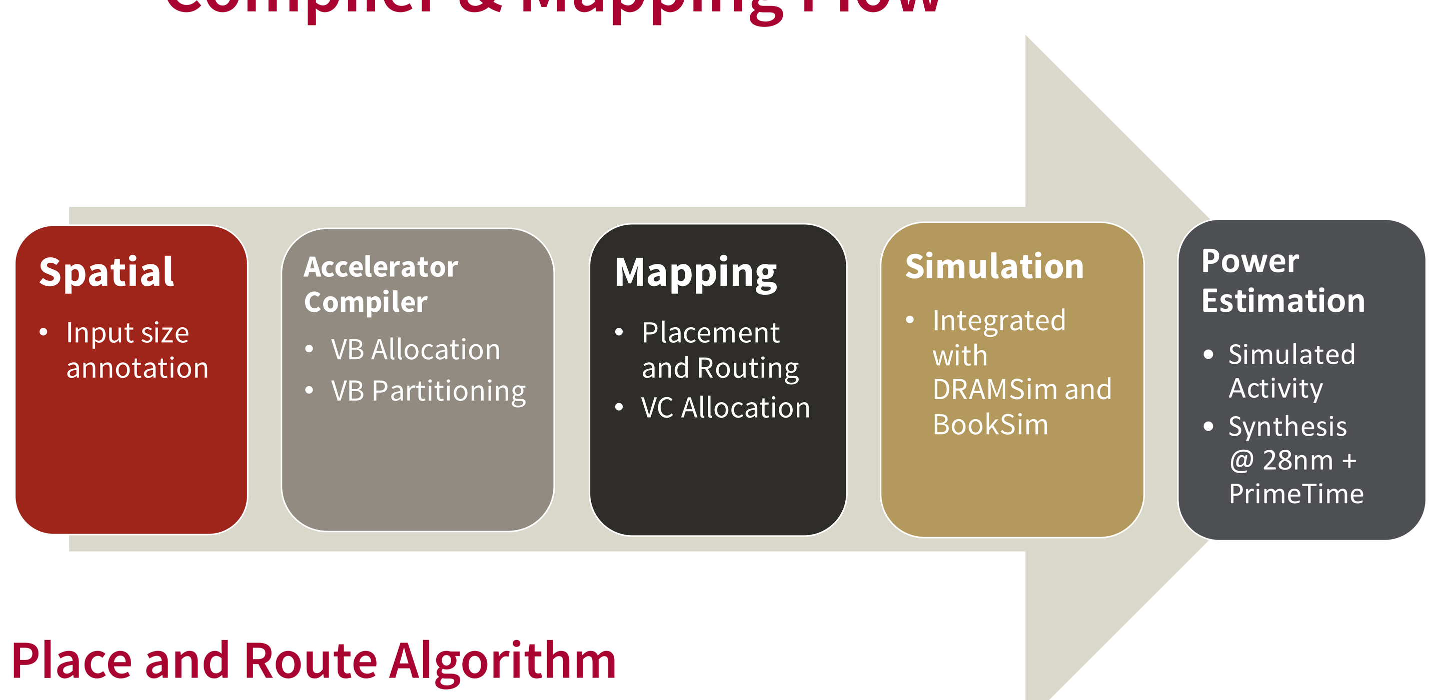
- Saturate compute throughput
- Low area and energy overheads
- Flexible for new applications
- Scalable to large arrays

Communication Pattern



Architecture	Comm. Type	Comm. Freq.	Granularity	Limited by
Processor	Parallelism	Low	Packet	Latency
Spatial Accelerator	Parallelism & Pipelining	High	Fine-grained	Throughput

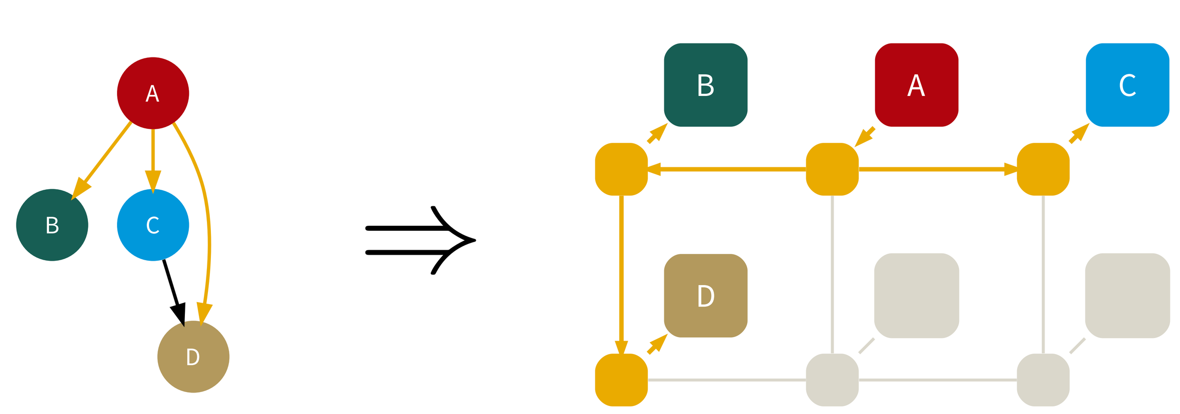
Compiler & Mapping Flow



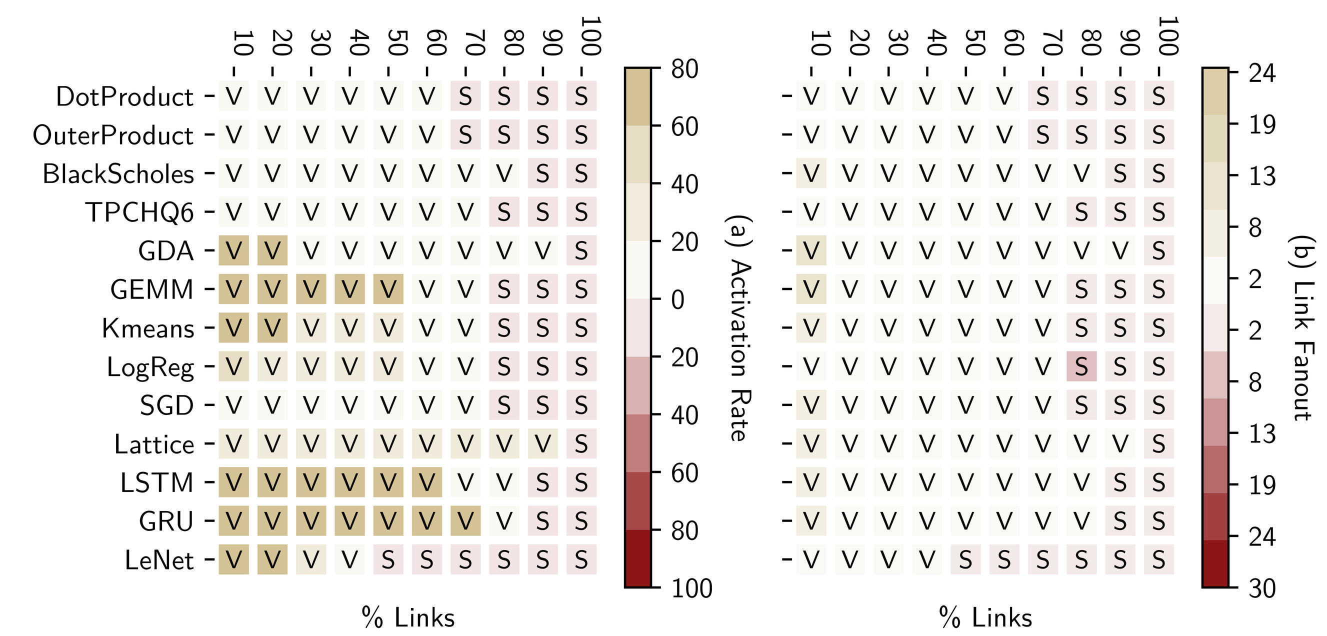
Place and Route Algorithm

- Start with random placement
  - Route all links, in order of activation count
    - Build most efficient broadcast tree
    - Guarantee static network placement, if possible
    - Else, map the link to the dynamic network
- Re-place VBs with the highest routing cost

$$Cost = f(DynCongest, avg(RouteLength), max(RouteLength))$$



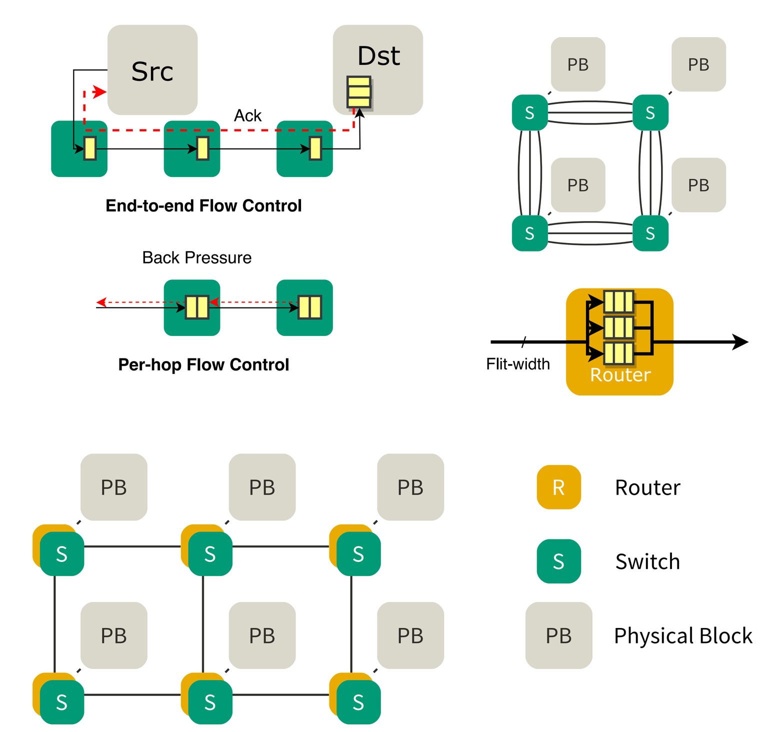
Observed Program Characteristics



How can we improve link usage and saturate compute/memory throughput?

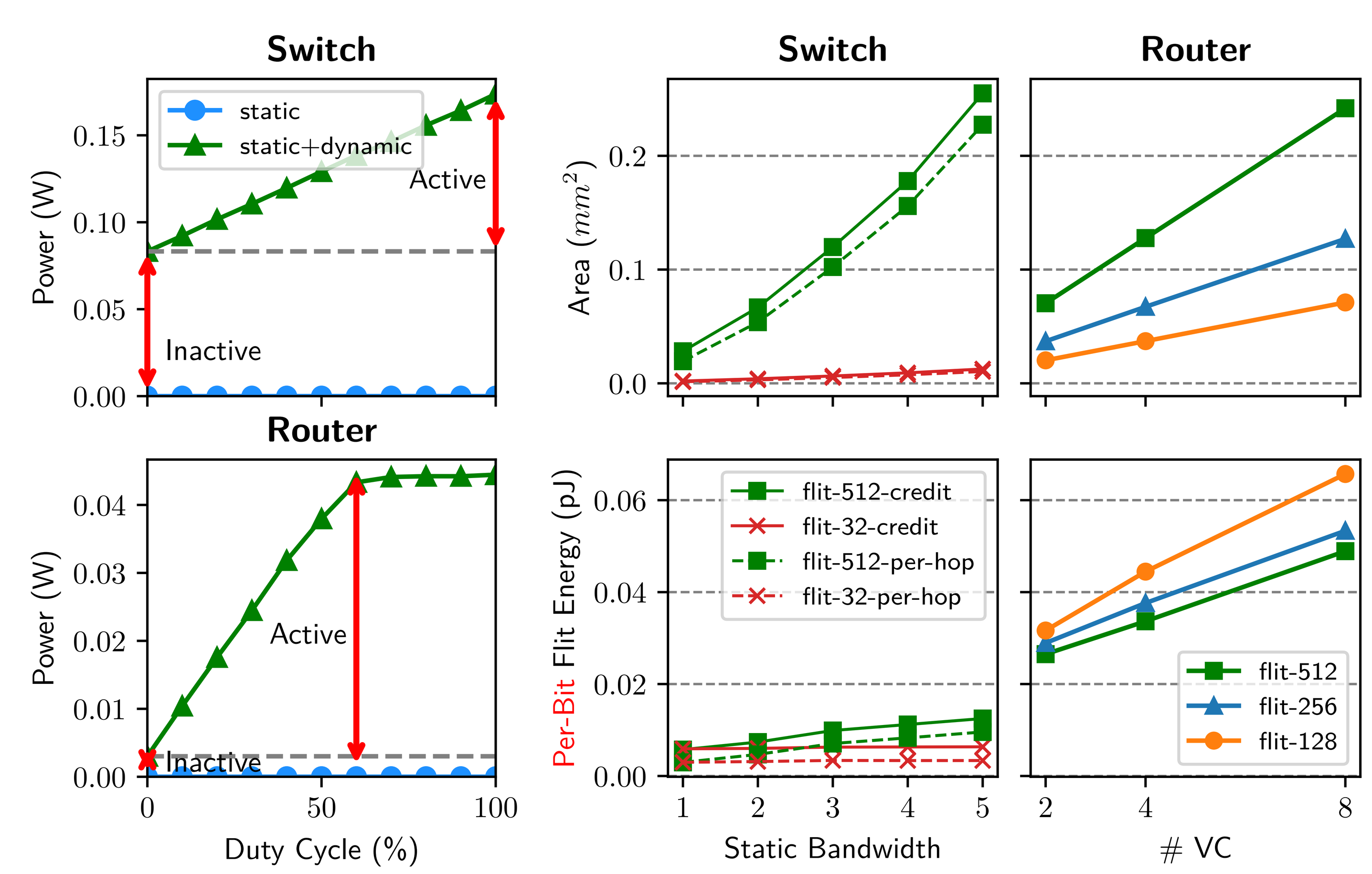
Network Design Space

- Static network
  - Flow control
    - End-to-end
    - Per-hop
  - Bandwidth
  - Scalar-only network
- Dynamic network
  - Virtual Channels (VCs)
  - Router flit width
- Static and dynamic hybrids
  - Varied static bandwidth
  - Varied dynamic params

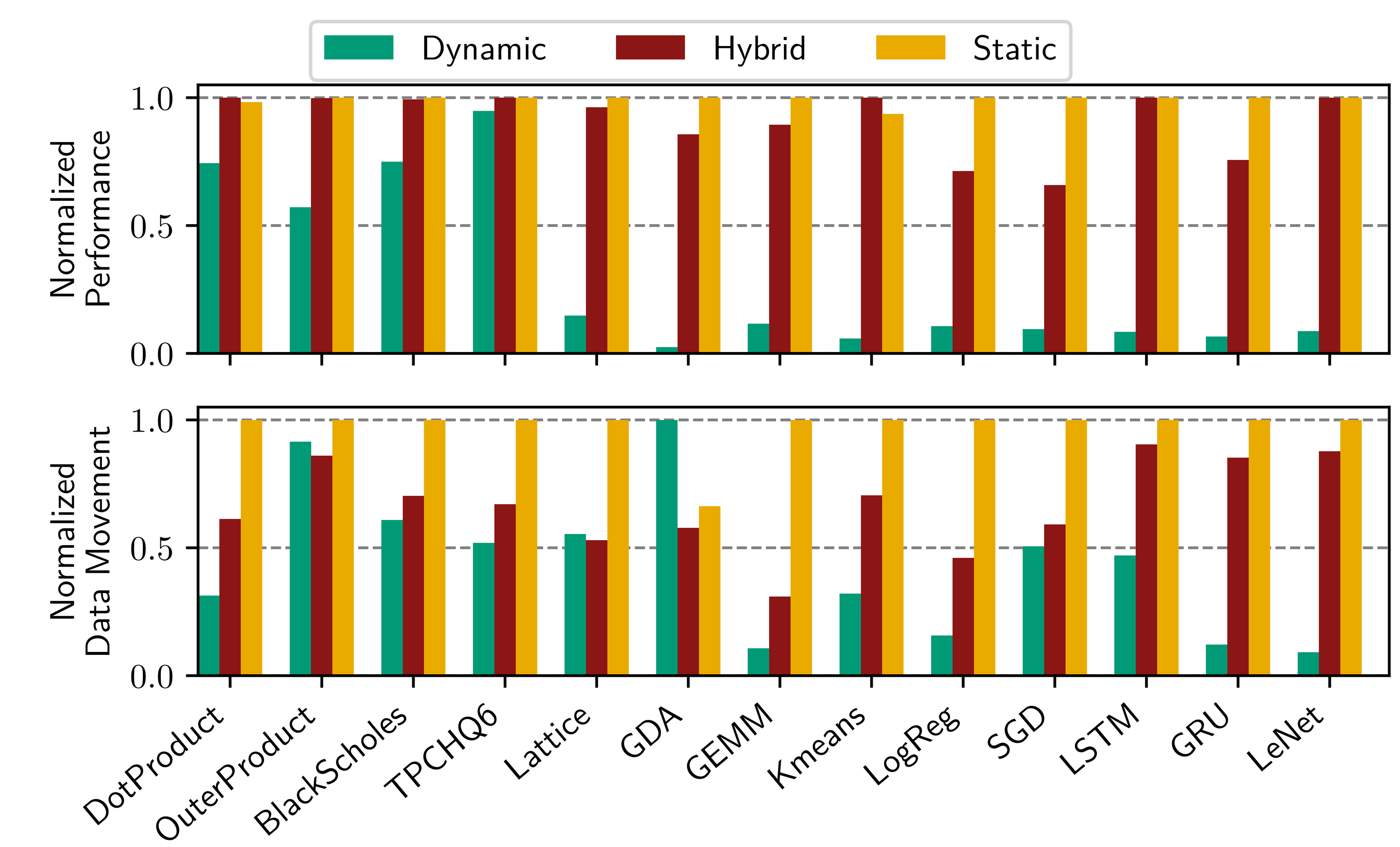


Area and Energy Characterization

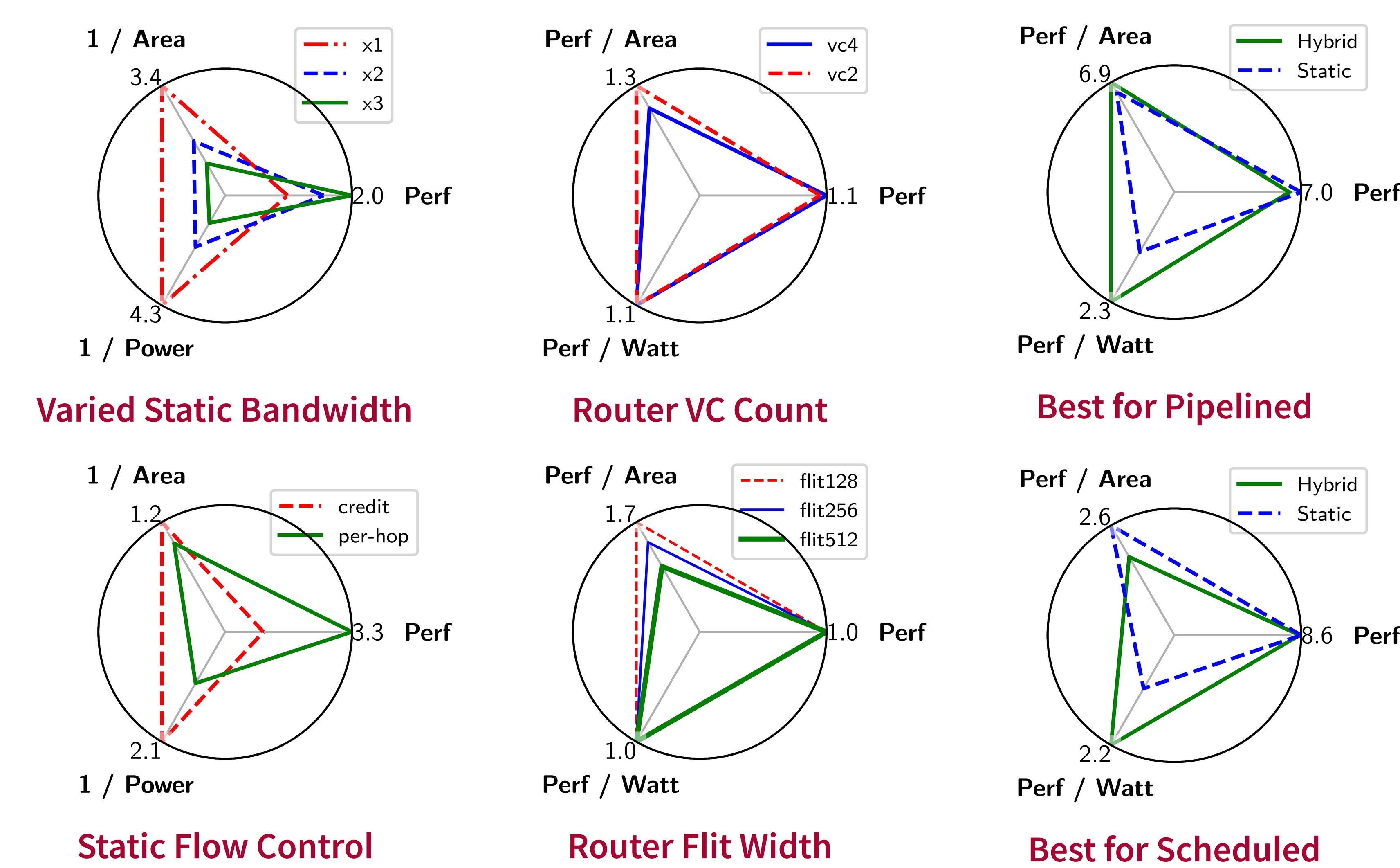
$$E_{net} = \sum_{\text{allocated}} P_{\text{inactive}} T_{\text{sim}} + E_{\text{flit}} \#\text{flit}$$



Evaluation



- Dynamic networks perform poorly on compute-bound applications due to low bandwidth
- Hybrid networks reduce data movement by using a dynamic network as an escape path



- On pipelined CGRAs, a hybrid network improves energy efficiency by ***1.8x*** compared to a static network with similar performance
- Performance varies up to ***7x*** between the best and worst network configurations

Conclusion

- Performance correlates strongly with network ***bandwidth*** for spatial accelerators
- Bandwidth scales more efficiently on a static network
- Combining large static and small dynamic networks:
  - Eliminates place and route failure
  - Improves ***perf/watt***