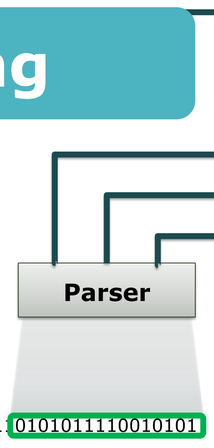
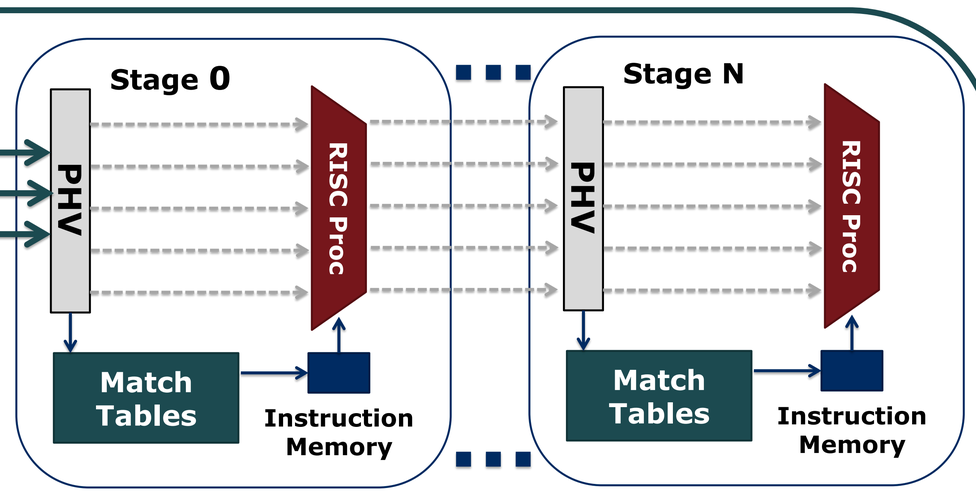
In-network Neural Networks

G. Siracusano, R. Bifulco
NEC Laboratories Europe, Systems and Machine Learning Group

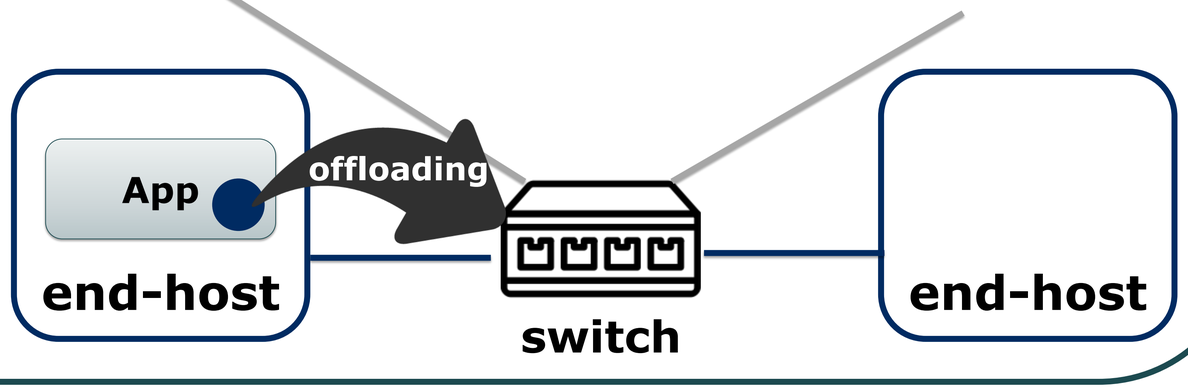
Background: In-network computing

- State-of-the-art **switching chips** are programmable
  - Programmable network packet parsing
  - Programmable packet modification instructions
  - Programming language e.g., **P4**





- **Performance**
  - Terabits/second (i.e., Billions of packets/second)
- **Limitations**
  - Simple arithmetic (+,-) /bitwise logic (AND, OR, …)
  - Small memory (10s MBs), small data bus (512B)
  - Small number of instructions per packet



A neural network on a network switch?

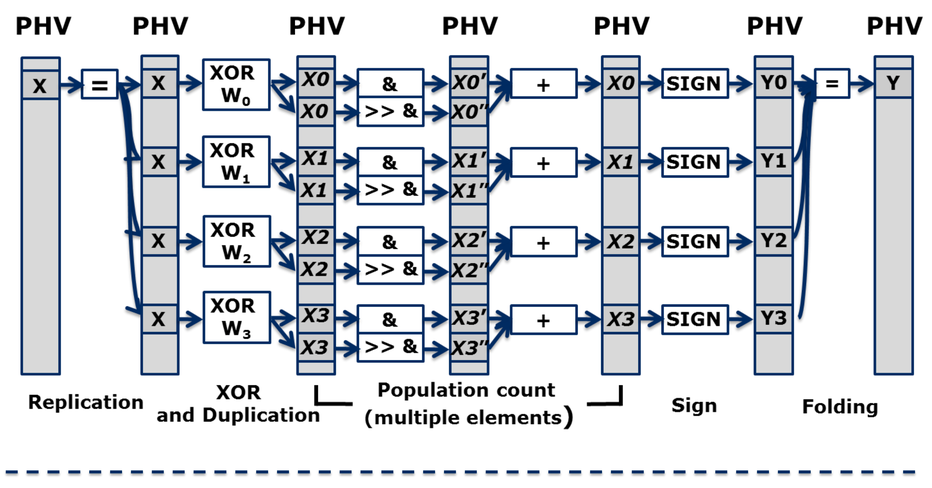
http://knowyourmeme.com/photos/859202-webcomics

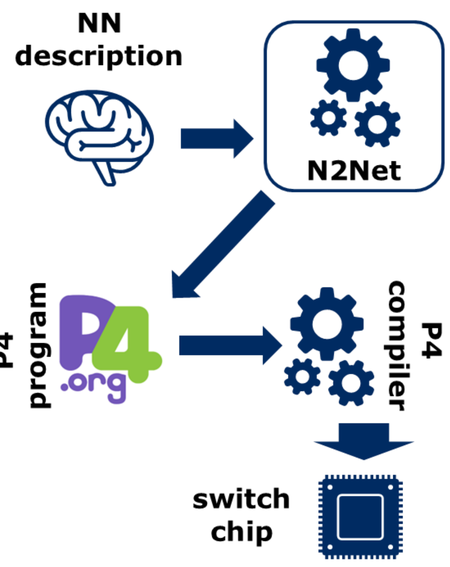
- Replace heuristic algorithms with specialized versions
  - Packet scheduling
  - Load balancing
  - Queue management
- Computation offloading
  - Pre/Post-processing
  - Informed "next-hop" selection
- Complement lookup tables for (packet) classification
  - White/black lists
  - Multi-stage classification
  - Specialized hash-functions
- Your use cases:

N2Net

- Regular Neural Networks
  - Multiplications ✗
  - Act. Func., e.g., Sigmoid ✗
- Binary Neural Networks
  - XNOR ✓
  - Popcount, Sign ✓

```
int popcount64a(uint64_t x)
{
    x = (x & m1 ) + ((x >>  1) & m1 );
    x = (x & m2 ) + ((x >>  2) & m2 );
    x = (x & m4 ) + ((x >>  4) & m4 );
    x = (x & m8 ) + ((x >>  8) & m8 );
    x = (x & m16) + ((x >> 16) & m16);
    x = (x & m32) + ((x >> 32) & m32);
    return x;
}
```





- With current hardware
  - line rate throughput with 96 (64, 32) neurons, 32b act.
  - 960M forward passes/sec
- Supporting larger networks at line rate:
  - Likely, ~+5% in chip's cost

Act. bits	32	64	128	256	512	1024	2048
Parallel neur.	64	32	16	8	4	2	1
Pipeline stages	14	16	18	20	22	24	25

NEC
Systems and Machine Learning Group
We are hiring!
Check open positions at: neclab.eu/jobs