InferX™ X1

High Performance Inference for Power Constrained Applications

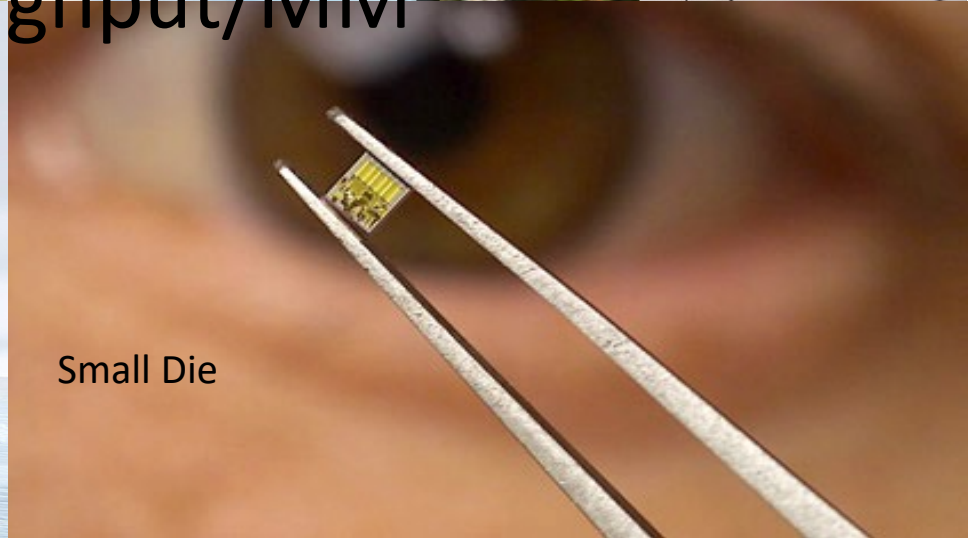
Cheng Wang Linley Spring Processor Conference, April 23 2021

High Throughput, Low Latency

High Accuracy

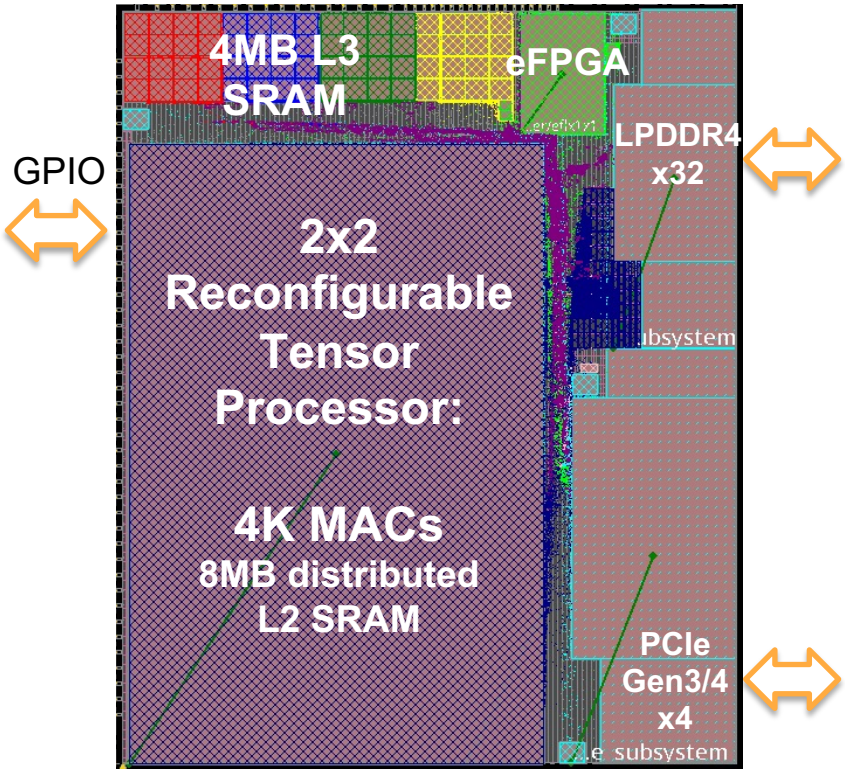
InferX X1 Higher Throughput/MM^2

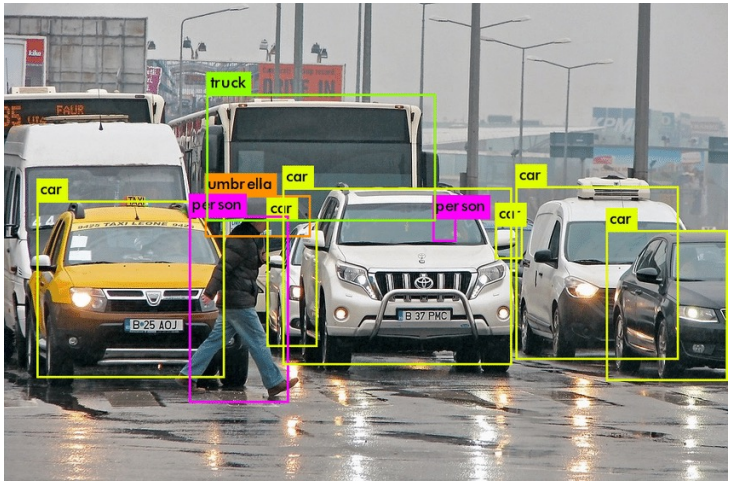
Large Models Megapixels Images



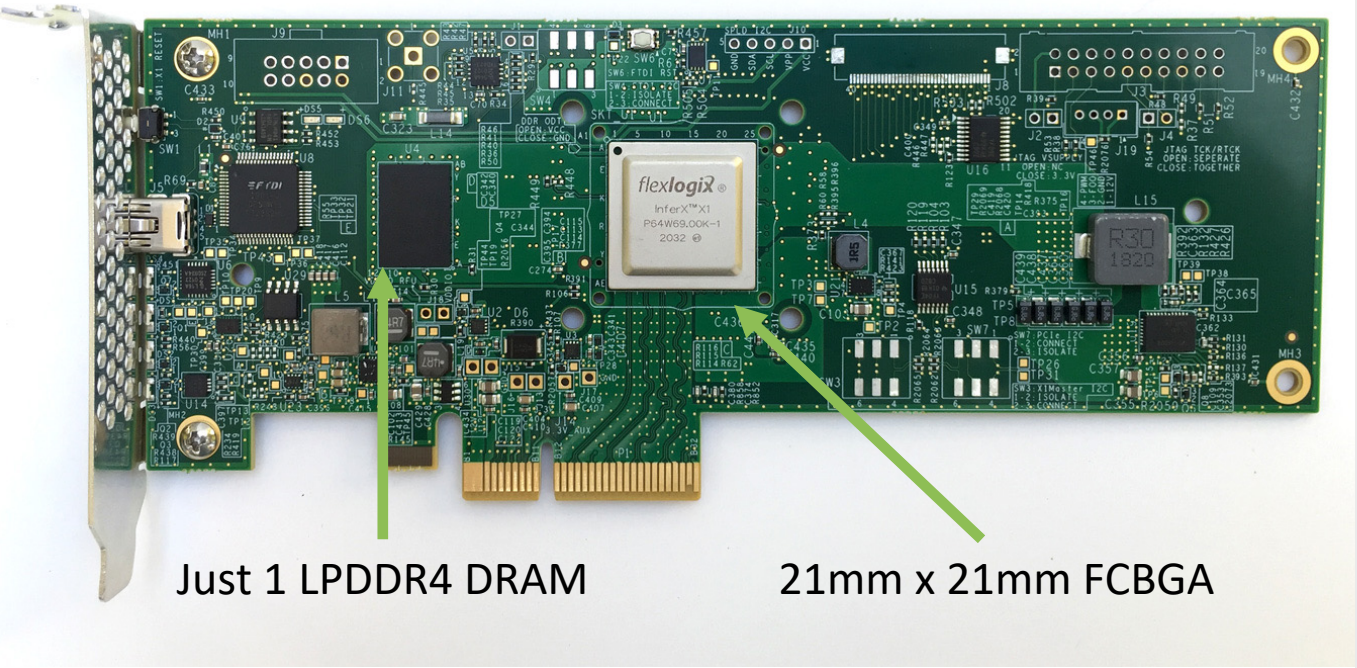
Small Die

InferX X1 Now Running YOLOv3





X1P1 – Performance at a fraction of $ and W



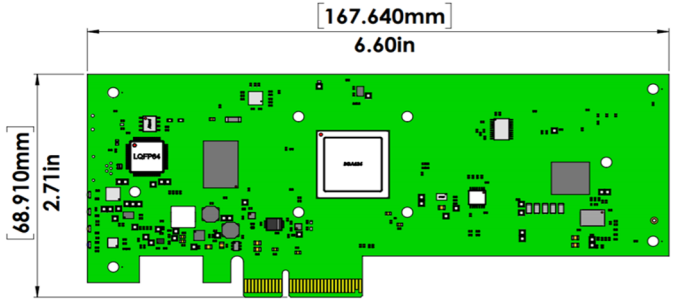
- 16W TDP Card
- Half Height
- Half Length
- x4 PCIe GEN3/4
- Available June

X1M: Smaller & Lower Power to put Performance Inference Anywhere

- M.2 form factor brings inference to mechanical and power constrained applications
- Provides faster path to production ramp vs custom card design

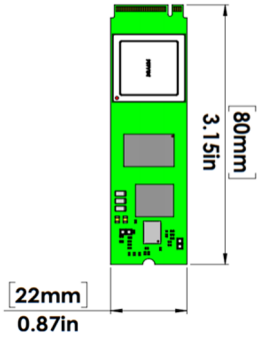
Size & Power Comparison

Half Height PCIe Card



16W TDP

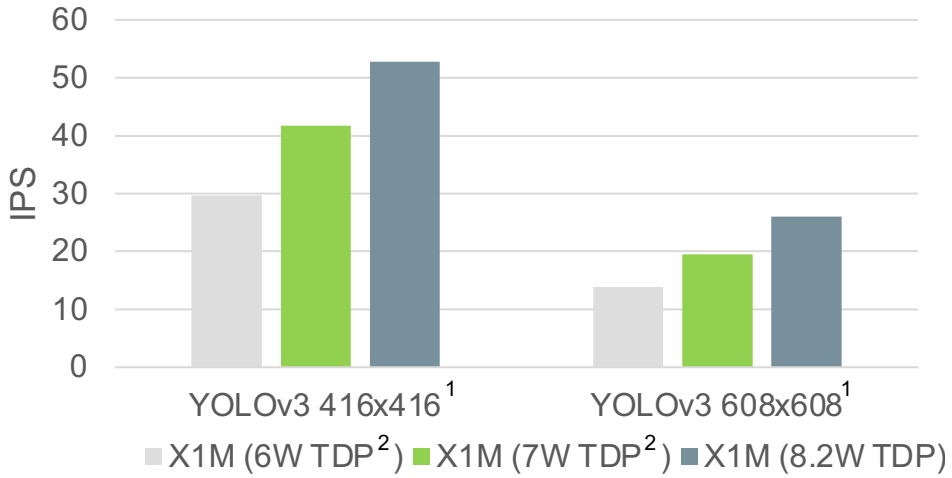
M.2 22X80 Card



8.2W TDP

X1M delivers high performance and high compute density

- More throughput for a given volume with only a 17.6 cm^2 footprint
- X1M can be throttled to fit different power / thermal requirements
  - More flexibility for customer applications
  - Even a 6W TDP X1M can run “heavy weight” models like YOLOv3 608x608 with <75ms latency

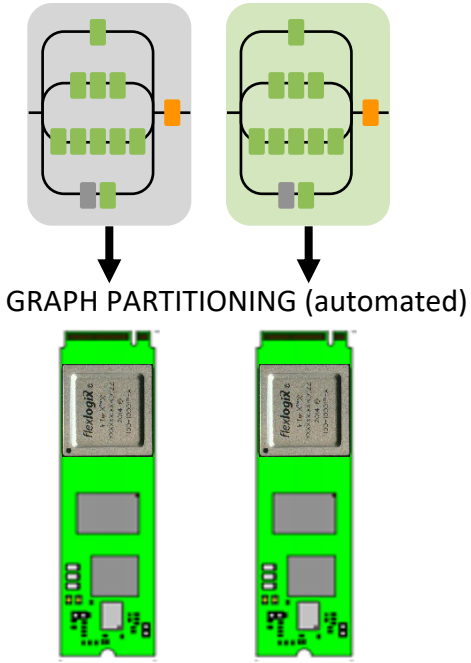


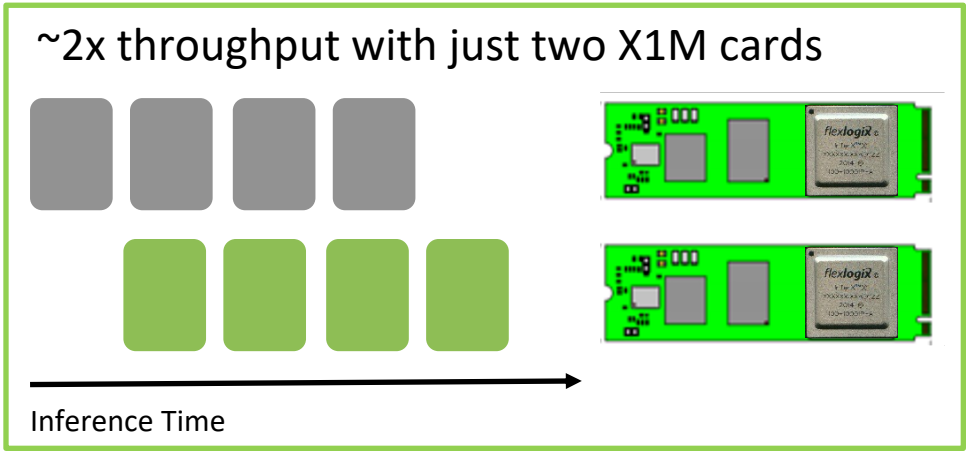
[1]: INT8, batch=1

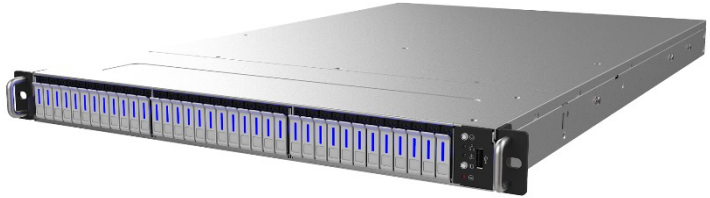
[2]: potentially further saving from lower VDD

Higher Compute Density = Higher Throughput for Edge Servers

- M.2 form factor allows for much more devices in each edge server
- Delivers superior inference throughput despite lower performance-per-card vs X1P1



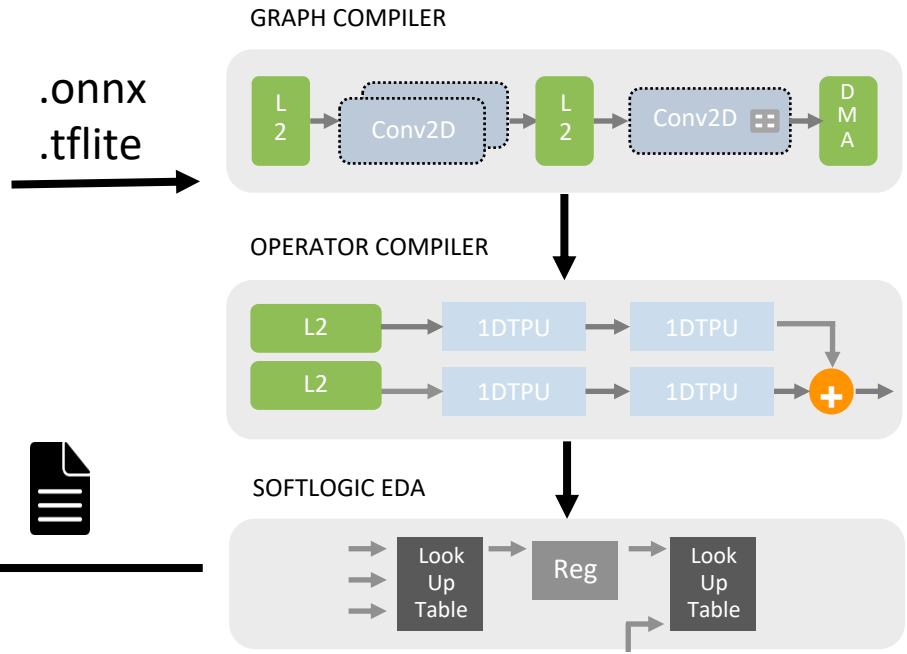




X1 Compiler Suite Optimizes IPS for Your Configuration

- X1 Compiler Suite accounts for # of X1 devices in your system configuration
- Optimizes graph partition inside Graph Compiler, such as:
  - Workload balancing across devices
  - Reducing chip-to-chip communication
  - Grouping similar layers on same device to enable configuration-reuse

- No manual partitioning or X1 Compiler experience is required!



Unified X1 Compiler Enables Your Development Now

Q2 2021

X1 Runtime API

- Support for Ubuntu, CentOS on precompiled .ncf models

Q3 2021

X1 Compiler Suite

- Support X1P or X1M configurations
- Support common INT8 operators (conv2d, add, concat, etc.)

Q4 2021

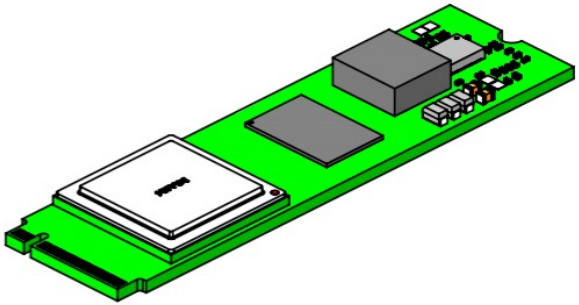
X1 Compiler Suite

- Support common BF16 operators

X1 Runtime API

- Support for Windows

InferX X1M M.2 Board Puts Inference Everywhere



InferX X1M 1KU Budgetary Pricing
$399

- 6-8.2W TDP X1M
- 22mm x 80mm
  B+M Key
- x4 PCIe GEN3/4
- Sampling 3Q21
- Production 4Q21

Let Us Benchmark Your Model

Do you have high resolution models that need to run in realtime?
We will show you how fast X1 can run them.

Dana McCarty
InferX Solutions
dana@flex-logix.com

TensorFlow Lite

ONNX