LSTM应用实例——实时数据序列


```
typedef mtl::LSTM<40, 10, 10, 4> RTLSTM;


RTLSTM::InMatrix<5> inMx;
RTLSTM::OutMatrix<1> outMx;
RTLSTM::OutMatrix<1> expectMx;


RTLSTM::InMatrix<10> inMx;
RTLSTM::OutMatrix<1> outMx;
RTLSTM::OutMatrix<1> expectMx;


RTLSTM::InMatrix<10> inMx;
RTLSTM::OutMatrix<3> outMx;
RTLSTM::OutMatrix<3> expectMx;
```

一个人能将某事教授给其他人，他算是真正地了解了这件事。一个程序员能将某件事教授给计算机他才算真正地了解了这件事。

未来计划和遗留问题

- 梯度消失和梯度爆炸
- 矩阵运算GPU支持
- 动态调整结点
- 大对象计算公式的运算符重载

$$A = B \times C$$

```
A.multiply(B, C);
```

结束

- 源码

  https://github.com/bowdar/DeepLearning

- 强烈推荐

  https://github.com/wichtounet/dll

- 参考资料

  http://www.asimovinstitute.org

  https://www.zybuluo.com/hanbingtao/note/433855

2017 CPP-Summit

深度学习框架与大规模深度学习训练系统
——来自商汤的解决方案

张行程

商汤科技深度学习框架核心开发者

SenseTime in AI Industry Landscape

Fundamental Algorithm Layer

- The technical barrier is high, the market scale is over 100 billion
- Have chance to generate a new giant company

Parrots | Tensor Flow | Torch

Hardware Layer

- capital intensive, not easy for startups to surpass

Application Layer

- Market Decentralized
- Computer vision technologies are the first to be monetized

Computer Vision

Voice Recognition

Autonomous Driving

IOT

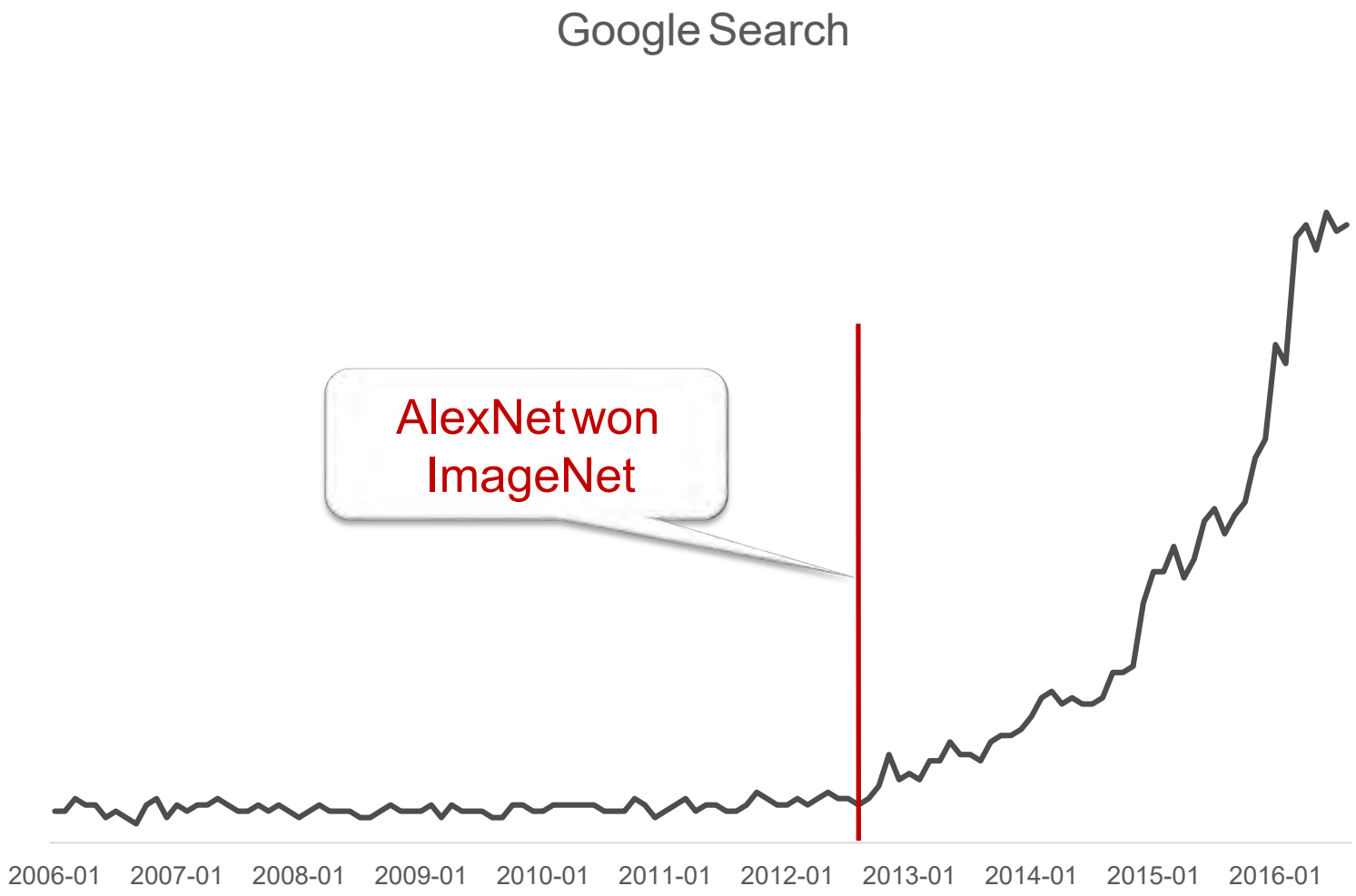
Healthcare

Cloud

Robotics

Business Intelligence

The Success of Deep Learning



Deep Learning Enables AI Breakthroughs

Facial Recognition

Voice Recognition

Image Recognition

Game Playing

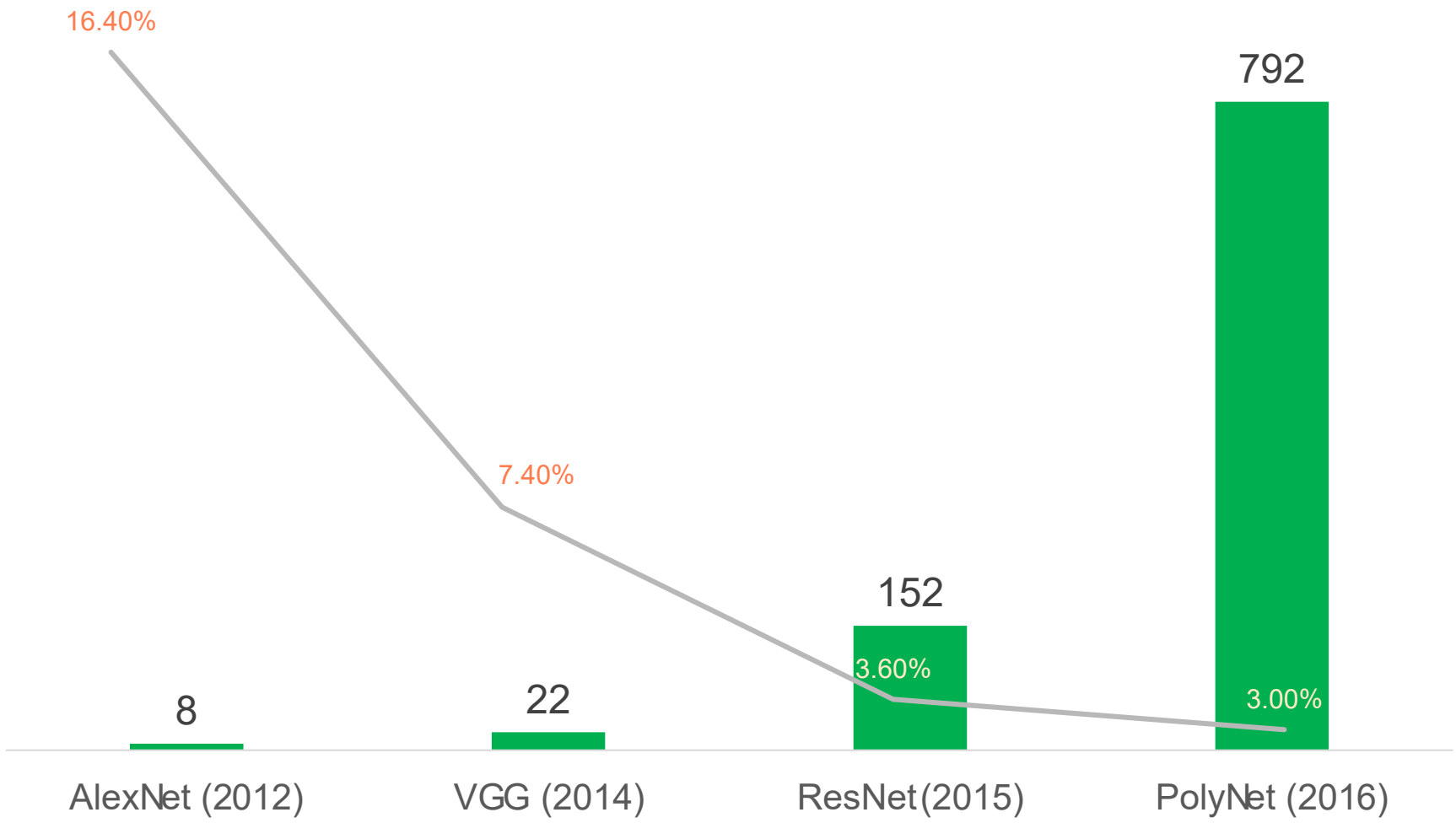
Deep Learning

Enabling machines to acquire knowledge and skills inducted from massive data

Autonomous Driving

What Lead to the Success?

The Race of Network Depth



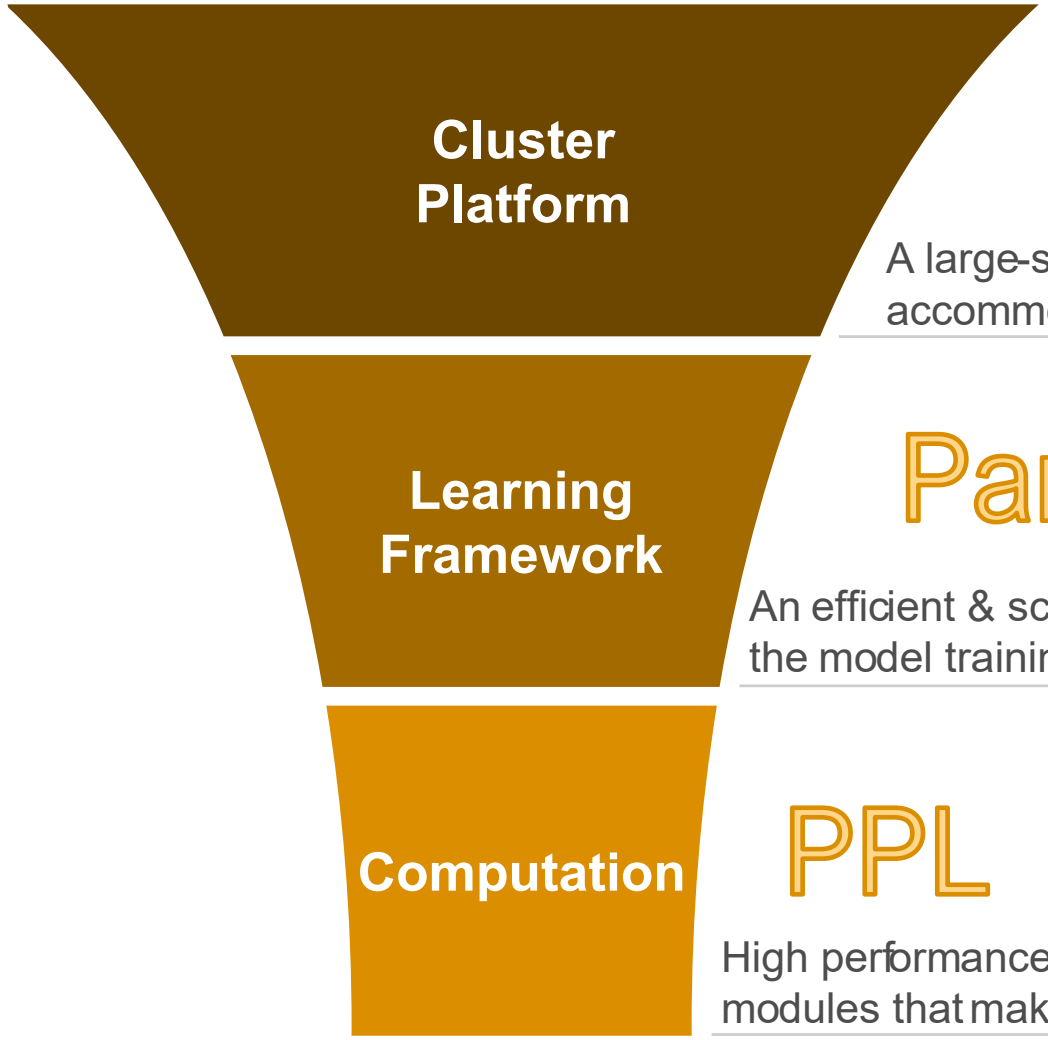
Challenges for training deep models

- **Dense computation requirement**
  - To train ResNet50: 1000,000 T FLOP
  - 14 days on NVIDIA M40 GPU
- **Limited GPU memory**
  - Largest memory available: Quadro M6000 24G
  - Requirement increase linear with model depth

More Challenges

- **Heavy concurrent small file access**
  - ~10T image data will be randomly accessed
  - Images size 10k~ 500k
  - Throughput should up to 2k images / sec
- **Large communication load**
  - Exchange 500 MB/s among every computation node

Technical Infrastructures



DeepLink

A large-scale cluster platform that can accommodate multiple concurrent tasks.

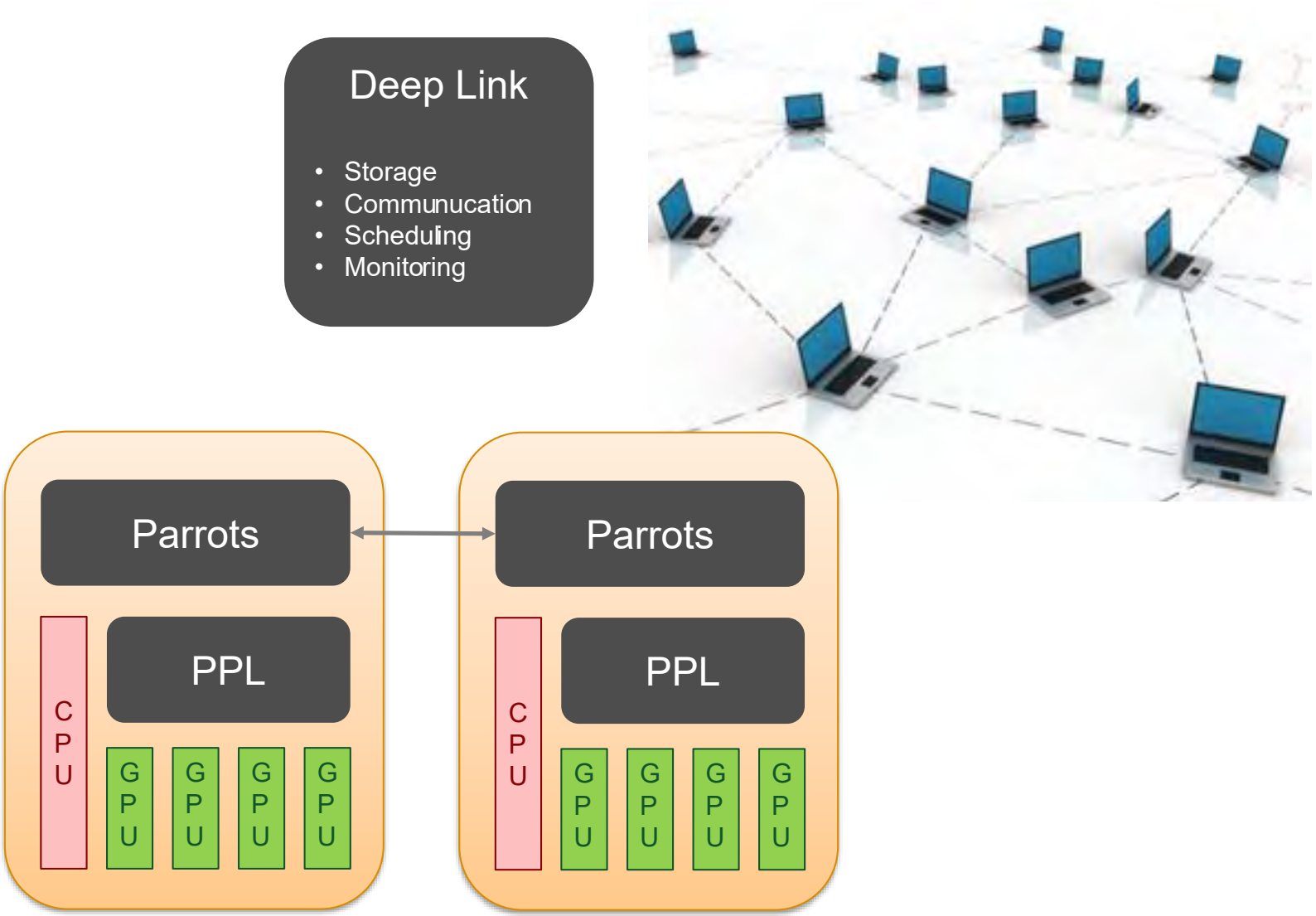
Parrots

An efficient & scalablelearning framework for performing the model training & testing.

PPL

High performance computational functions or hardware modules that make each step fast

Our Integrated Solutions



PPL

A library of highly optimized computational modules.

1

Parrots

A deep learning framework that is **efficient**, **scalable**, and **flexible**

2

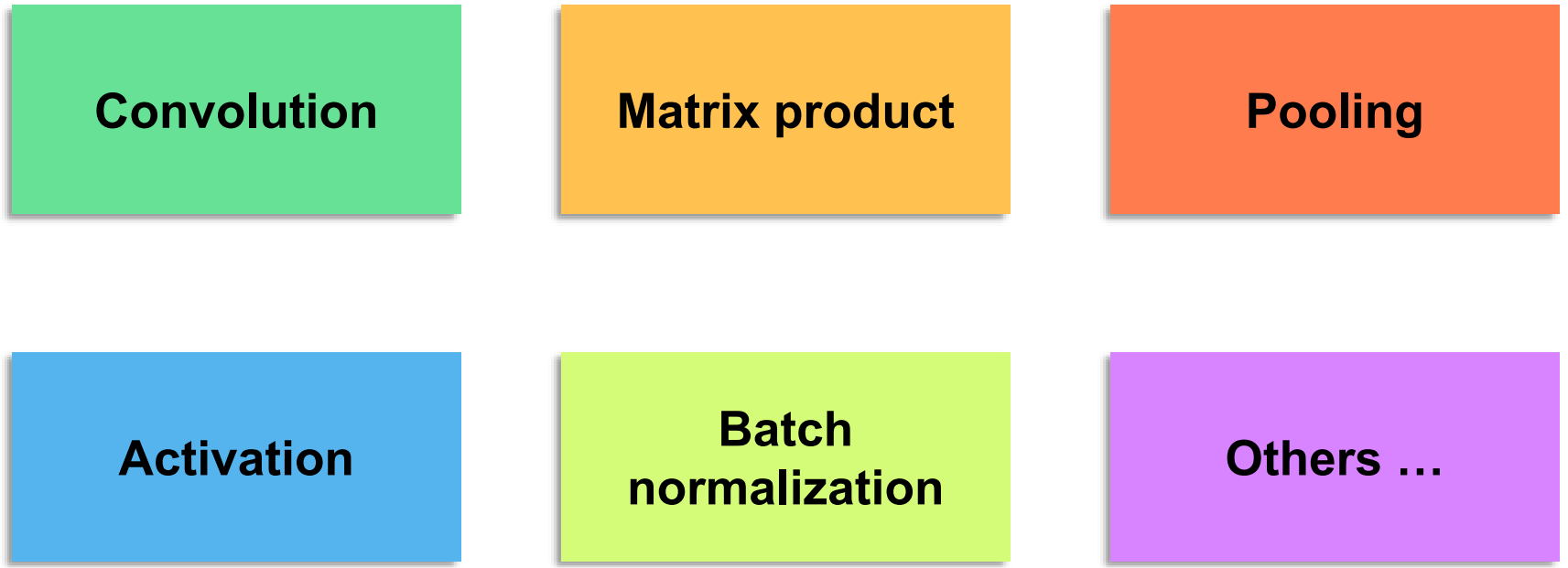
DeepLink

A large-scale cluster platform designed for deep learning.

3

Layers

The efficiency of a deep network essentially depends on how fast we can compute:



PPL (Parrots Primitive Library) is to provide highly optimized functions for such computation.

PPL Editions

PPL has multiple editions for different architectures

X86

- Support all common CNN functions.
- Support Windows, Linux, and Mac OS X.
- Support ISA: AVX, AVX2, and 64-bit processors.

ARM

- Support all common CNN functions.
- Support iOS, Android, Linux, and Windows.
- Supported ISA: ARMv7 (32bit), ARMv8 (64bit)
- Supported processors:
  - ARM Cortex-A
  - iPhone 4x – 6x
  - All qualcomm

CUDA

- Support all common CNN functions.
- Support Windows, Linux, and Mac OS X.
- Support all NVidia GPUs
  - GeForce
  - Tesla
  - Tk1
- cuDNN + our own implementations

OpenCL

- Mainly to support embedded devices.
- Support all common CNN functions.
- Supported devices:
  - Qualcomm adreno GPU
  - ARM Mali GPU
- Performance comparable to the ARM edition.

PPL X86 vs. Intel MKL