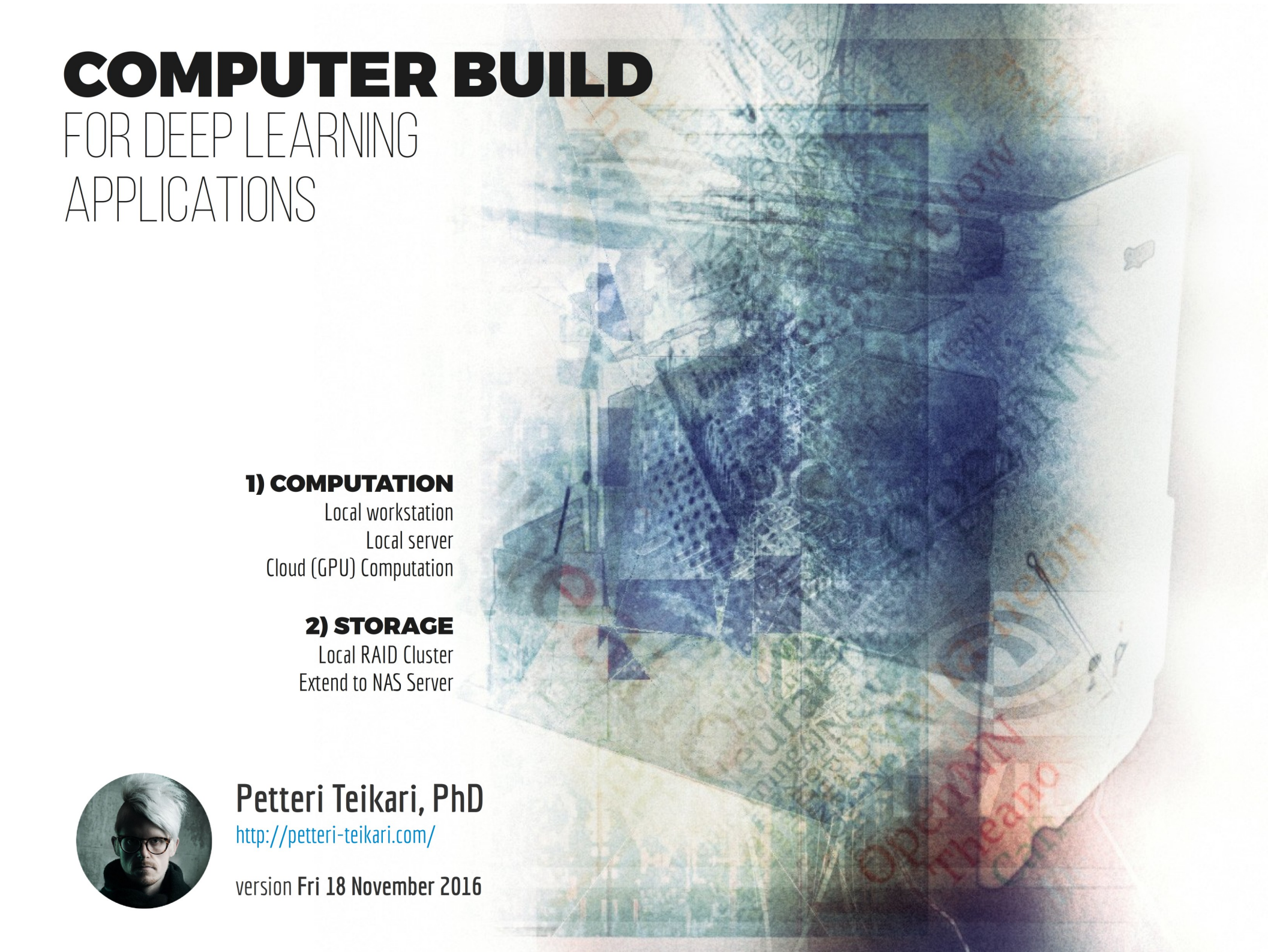
COMPUTER BUILD

FOR DEEP LEARNING APPLICATIONS



1) COMPUTATION

Local workstation
Local server
Cloud (GPU) Computation

2) STORAGE

Local RAID Cluster
Extend to NAS Server

Petteri Teikari, PhD
http://petteri-teikari.com/

version **Fri 18 November 2016**

COMPUTATION WORKSTATION

You cannot get the best performing system with the least amount of money

→ compromises have to be made

→ no 'one-size-fits-all' solution (what do **you** need?)

PROPOSITION

Start off with a 'server skeleton' that can house 4 or 8 GPUs (Titan X) and handle the possible peak training demands with cloud services

NOTE 1!

This information can outdate quickly due to the rapid advances in the 'deep learning computing'

NOTE2!

The proposed system is for academic / small startup budgets, for personal/multi-use (gaming, rendering, etc.) one can check the e.g. Andrej Karpathy's ~$3,000 system

How to read the slideshow

- Meant as an **'introductory' deck** for all the components typically needed when building your own workstation with some storage.
  - Main focus demographic for the slideshow is **university labs**, and **small startups** which do not have huge budgets (under £15,000) but still more money than individual 'dorm room Kagglers'
  - Hard to provide complete package for everyone with individual needs.
- Slides are quite condensed and best read with tablet with an easy zooming.
  - Intent to provide starting points for multiple aspects of the system that you need to consider rather than trying to cover everything in detail
- Prices are indicative, and there might be variations of the same product within slides as they are snapshot at different times.
- If you find **errors** on this, I would appreciate bug reports :)

System Specifications 2D/3D IMAGE PROCESSING

- Dual CPU system based on Intel Xeon CPUs
  - 4 x PCIe 16-lane GPUs achieved with dual CPU
- 2-8 x Pascal NVIDIA Titan X (12 GB)
  - Maximum GPU RAM with great performance as well.
    Note! Not all frameworks necessarily can use multiple GPUs
- 256 GB RAM
  - Faster to develop for non-GPU pre-processing / data wrangling especially for large 3D datasets without having to worry too much of running out of memory. Cache data also for GPU.
- 3-tier HDD setup
  - fast M.2 SSD cache, SATA SSD for OS, Spinning HDDs for storage

FOR BACKGROUND AND ALTERNATIVES FOR THESE CHOICES, SEE FOLLOWING SLIDES

Options Available OVERVIEW

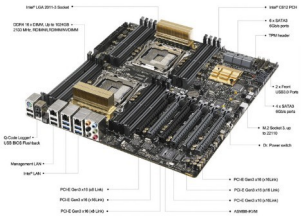
- For our needs, there are 4 main lines that we could base our setup. The choice depend on your current and future needs.

SINGLE CPU INTEL I7 "CONSUMER", **4XGPU**

X99-E WS

i7-based systems
Limited to 64 GB
of memory

DUAL CPU INTEL XEON "CONSUMER", **3XGPU**



Z10PE-D16 WS

"Consumer" dual
Xeon MB with
4xGPU slots?

Similar upfront cost

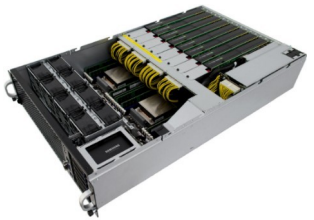
This slideshow focuses on **Intel+NVIDIA** solutions due to their "established status". It would be good to have smaller players for sure challenging the gorillas of the field.

DUAL CPU INTEL XEON "SERVER", **4XGPU**

SC747TQ-R1620B

Slots for 4 GPUs,
and still room for a
network card.

DUAL CPU INTEL XEON "SERVER", **8XGPU**



Asus ESC8000 G3

Slots for 8 GPUs,
and still room for a
network card.

Higher upfront cost than "consumer" solutions

Good for GPU clusters, and for growing organizations

Commercial workstation EXAMPLE #1

Not totally ridiculous markup for the similar components chosen

Chillblast Fusion DEVBOX Deep Learning PC, https://www.chillblast.com/chillblast-fusion-devbox-deep-learning-pc.html

Commercial workstation EXAMPLE #1

microway.com

- 6-core Xeon “Haswell” processor (12 threads)
- 64GB DDR4 memory,
- 4 x NVIDIA GeForce GTX Titan X GPUs,
- 1 x SSD for the operating system,
- 6TB of high-speed RAID10 storage

WhisperStation™- Deep Learning

Ultra-Quiet Computing for Deep Learning Researchers

Designed for deep learning research, this WhisperStation features a high-speed 6-core Xeon “Haswell” processor, 64GB DDR4 memory, four NVIDIA GeForce GTX Titan X GPUs, an SSD for the operating system, 6TB of high-speed RAID10 storage (with SSD caching, if desired) and ultra-quiet fans. We provide a turn-key system delivered with Ubuntu Linux, NVIDIA CUDA 7.5, cuDNN v3 and your choice of machine learning software pre-installed (e.g., Caffe, Torch, Theano, NVIDIA cuDNN, NVIDIA DIGITS).

The *WhisperStation-Deep Learning* workstation is a commercially-available system based upon NVIDIA’s DIGITS DevBox. It has been designed to provide maximum performance when training deep neural networks. It is the perfect system for a researcher exploring the applications of machine learning before such applications move into production. Using NVIDIA’s DIGITS interface for popular deep learning tools, such as Caffe, researchers can collaborate and share GPU resources as they fine-tune their neural networks.

For production-level deep learning, we recommend our professional NumberSmasher server product line with NVIDIA Tesla GPUs.

Features | Specifications | Accessories/Options | Support | Price

System Price: $9,487+

Each Microway system is customized to your requirements. Final pricing depends upon configuration and any applicable educational or government discounts.

CONFIGURE YOUR SOLUTION
DESIGN A HARDWARE CONFIGURATION THAT MEETS YOUR NEEDS.

TALK TO AN EXPERT
CHALLENGE US WITH YOUR COMPUTING PROBLEM.

CUSTOMER TESTIMONIALS

I really appreciate all your effort, we have dealt with several other vendors and you have gone well above and beyond what they offered.

» Daniel K.
Manufacturing Automation

Commercial server EXAMPLE

Home / Featured Systems / 40 Tflops 3U rackmount using eight Tesla K40s

40 Tflops 3U rackmount using eight Tesla K40s

Built by Oscar on 03/26/15

Share: 2 49 100

Full Specs

Asus System ESC8000 G3 3U Server
2x Intel Xeon E5-2660 V3 2.6GHz Ten Core 25MB 105W
8x Crucial DDR4-2133 16GB ECC Reg. (128GB Total)
Onboard Video
Samsung 850 Pro 1TB SATA 6Gb/s 2.5inch SSD (Primary)
4x Samsung 850 Pro 1TB SATA 6Gb/s 2.5inch SSD (Storage)
Ubuntu 14.04 LTS Installation w/ CD (64-bit)
8x NVIDIA Tesla K40 PCI-E 12GB (Passive)

www.pugetsystems.com/

TESLA GPUS FOR WORKSTATIONS

Feature	Tesla K40	Tesla K20
Peak double precision floating point performance	1.43 Tflops	1.17 Tflops
Peak single precision floating point performance	4.29 Tflops	3.52 Tflops
Memory bandwidth (ECC off)	288 GB/sec	208 GB/sec
Memory size (GDDR5)	12 GB	5 GB
CUDA cores	2880	2496

TITAN X

11 Tflops
480 GB/s
12 GB
3584

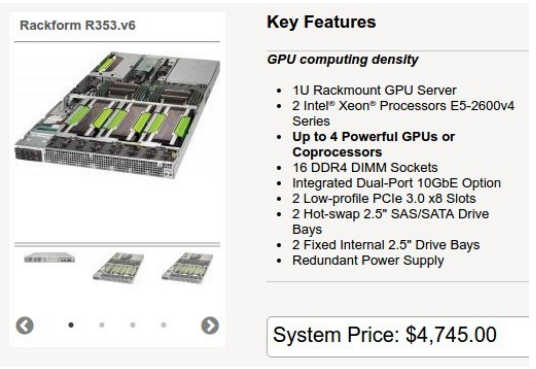
8 x 11 Tflops = 88 Tflops with 8 Titan X GPUs

Facebook's "Big Sur" machine learning server supports up to eight GPUs. Source: Facebook

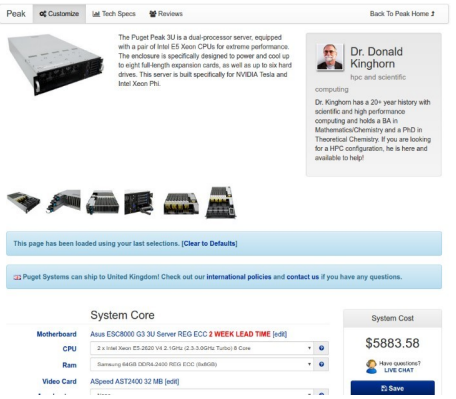
CADE METZ BUSINESS 12.10.15 12:00 PM

FACEBOOK OPEN SOURCES ITS AI HARDWARE AS IT RACES GOOGLE

8 x 7 Tflops (FP32)
= 56 Tflops with 8 Tesla M40 GPUs



siliconmechanics.com

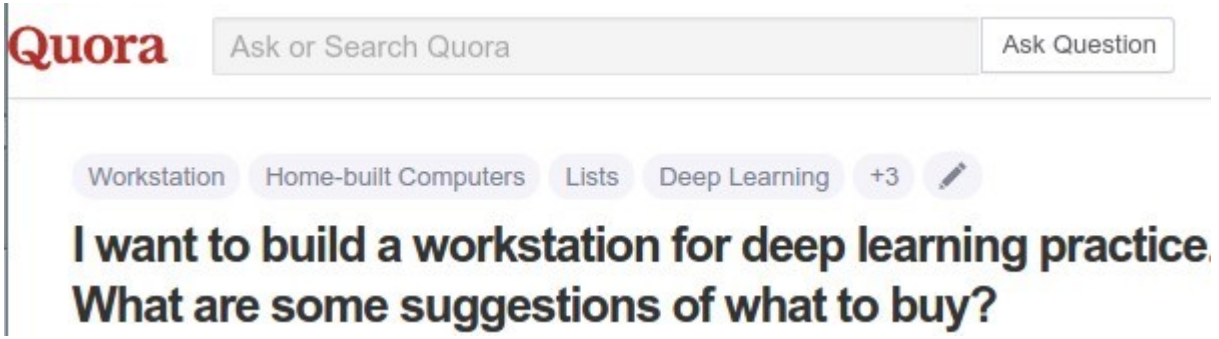


pugetsystems.com | Customize

Build your own Workstation

- Commercial solutions on previous slides really do not have any proprietary technology, and they have been built from off-the-shelf components.
- **Advantages** of building your own setup
  - Cheaper of course than buying a commercial setup
  - Flexibility of choosing the components that you really need
  - Getting to know your own system
- Some practical **problems** that you may face:
  - Most of the discussion online is focused on building gaming setups and the advice from them might not be that relevant for machine learning use.
  - Unsurprising incompatibility issues with different hardware (e.g. using Intel Xeon Phi accelerators)

Workstation Tips



quora.com

"If you want a high end PC, it is always economical to build. This holds true for machines with GPUs such as high end gaming systems or deep learning systems. It will always be worth the hassle, financially to order and build a machine unless your time is premium. As a example, the NVIDIA® DIGITS™ DevBox costs about 15,000$ with four nvidia GPUs. This is an ***outrageously* priced** system unless the box ships with an intern from NVIDIA. You can easily build a state of the art high end deep learning machine with two NVIDIA GPUs for just 2.5-4K$ each"

If you want to build a new machine, I put together an example configuration as of today (late 2015). I believe this is future proof until late 2016/early 2017 or when Skylake-E is announced:

a) An Intel 40 lane CPU (The latest Haswell-E processor Intel 5930K is the most value for money CPU with 40 PCI-E lanes). The only other Haswell CPU that supports 40 lanes is Intel 5960X but it is extremely expensive for the value it brings.

"The rule of thumb in building/buying PCs is if you need a commodity PC for programming/web-browsing (usually sub $1000), it is cheaper to buy. If you want a **high end PC**, it is **always economical to build**."

Workstation #2 TIM DETTMERS

- Excellent article by Tim Dettmers, do not forget to read.

A Full Hardware Guide to Deep Learning

2015-03-09 by Tim Dettmers — 278 Comments

"As we shall see later, **CPU cache** size is rather irrelevant further along the CPU-GPU-pipeline, but I included a short analysis section anyway so that we make sure that every possible bottleneck is considered along this pipeline and so that we can get a thorough understanding of the overall process.

In deep learning, the same memory is read repeatedly for **every mini-batch** before it is sent to the GPU (the memory is just overwritten), but it depends on the mini-batch size if its memory can be stored in the cache. For a mini-batch size of 128, we have 0.4MB and 1.5 MB for MNIST and CIFAR, respectively, which will fit into most CPU caches; for ImageNet, we have more than 85 MB for a mini-batch, which is much too large even for the largest cache (L3 caches are limited to a few MB).

Because data sets in general are **too large** to fit into the cache, new data need to be read from the RAM for each new mini-batch – so there will be a **constant need to access the RAM** either way.

http://timdettmers.com/2015/03/09/deep-learning-hardware-guide/

Before the end: how to select GPU for research

Select the right GPU can save you lots of money and boost your performance, you should read through this article before your purchase: `https://timdettmers.wordpress.com/2014/08/14/which-gpu-for-deep-learning/`
From conclusion of the article:

- best GPU overall: GTX Titan X
- cost efficient but still expensive: GTX Titan X or GTX 980
- cheapest card with no troubles: GTX 960 4GB or GTX 680
- I work with data sets > 250GB: GTX Titan
- I have no money: GTX 680 3GB
- I do Kaggle: GTX 980 or GTX 960 4GB
- I am a researcher: 1-4x GTX 980
- I am a researcher with data sets > 250GB: 1-4x GTX Titan

Conclusion / TL;DR

GPU: GTX 680 or GTX 960 (no money); GTX 980 (best performance); GTX Titan (if you need memory); GTX 970 (no convolutional nets)

CPU: Two threads per GPU; full 40 PCIe lanes and correct PCIe spec (same as your motherboard); > 2GHz; cache does not matter;

RAM: Use asynchronous mini-batch allocation; clock rate and timings do not matter; buy at least as much CPU RAM as you have GPU RAM;

Hard drive/SSD: Use asynchronous batch-file reads and compress your data if you have image or sound data; a hard drive will be fine unless you work with 32 bit floating point data sets with large input dimensions

PSU: Add up watts of GPUs + CPU + (100-300) for required power; get high efficiency rating if you use large conv nets; make sure it has enough PCIe connectors (6+8pins) and watts for your (future) GPUs

Cooling: Set coolbits flag in your config if you run a single GPU; otherwise flashing BIOS for increased fan speeds is easiest and cheapest; use water cooling for multiple GPUs and/or when you need to keep down the noise (you work with other people in the same room)

Motherboard: Get PCIe 3.0 and as many slots as you need for your (future) GPUs (one GPU takes two slots; max 4 GPUs per system)

Monitors: If you want to upgrade your system to be more productive, it might make more sense to buy an additional monitor rather than upgrading your GPU

Workstation #3a ROELOF PIETERS

I: Building a Deep Learning (Dream) Machine

> *As a PhD student in Deep Learning, as well as running my own consultancy, building machine learning products for clients I'm used to working in the cloud and will keep doing so for production-oriented systems/algorithms. There are however huge drawbacks to cloud-based systems for more research oriented tasks where you mainly want to try out various algorithms and architectures, to iterate and move fast. To make this possible I decided to custom design and build my own system specifically tailored for Deep Learning, stacked full with GPUs. This turned out both more easy and more difficult than I imagined. In what follows I will share my "adventure" with you. I hope it will be useful for both novel and established Deep Learning practitioners.*

http://graphific.github.io/posts/building-a-deep-learning-dream-machine/

So what is one to do? Simple: **get** your own GPU rig!

OVERVIEW

In short:

- Double precision (as Nvidia's Tesla K20/40/80 offer) is a waste of money as this type of precision is not needed for DNNs;
- Get a **motherboard** which supports PCIe 3.0 and supports PCIe power connectors of 8pin + 6pin with one cable, so you can add up to 4 GPUs. The motherboard should be able to support your GPU configuration, ie enough physicial lanes to support a x8/x8/x8/x8 setup for 4 GPUs;
- Get twice the amount of **RAM** as your total GPU memory;
- As for the **Power Supply Unit (PSU)** get one with as high efficiency as you can afford to, and take into account the total wattage you might need - again - now and in the future: Titanium or platinum quality PSUs are worth the money: you will save money and the environment, and get back the extra $$ in no time on saved energy costs. 1500 to 1600 Watt is what you probably need for a 4 GPU system;
- **Cooling** is super important, as it affects both performance and noise. You want to keep the temperature of a GPU at all times below 80 degrees. Anything higher will make the unit lower its voltage and take a hit in performance. Furthermore, too hot temperatures will wear out your GPU; Something you'd probably want to avoid. As for cooling there are two main options: Air cooling (Fans), or Water cooling (pipes):
- Get a chassis with enough space for everything. Bigger chassis offer more airflow. Make sure there are enough **PCIe slots** to support all the GPUs, as well as possibly any other PCIE cards you might install (as fast Gigabit network cards or whatever). One GPUs typically takes the space of 2 PCIe slots. In a typical chassis this means 7 PCIe slots, as the last GPU can be mounted at the bottom using only one slot;

Workstation #3b ROELOF PIETERS

In the end, after thorough reading, helpful replies from Tim Dettmers, and also going over Nvidia's DevBox and Gamer Forums, the components I chose to put together. It is clear that the machine is partly (at least the chassis is) inspired by Nvidia's DevBox, but for almost 1/2 of the price.

- Chassis: Carbide Air 540 High Airflow ATX Cube
- Motherboard: Asus X99-E WS workstation class motherboard with **4-way PCI-E Gen3 x16 support**
- RAM: 64GB DDR4 Kingston 2133Mhz (8x8GB)
- CPU: Intel(Haswell-e) Core i7 5930K (6 Core 3.5GHz)
- GPUs: 3 x NVIDIA GTX TITAN-X 12GB
- HDD: 3 X 3TB WD Red in RAID5 configuration
- SSD: 2 X 500GB SSD Samsung EVO 850
- PSU: Corsair AX1500i (1500Watt) 80 Plus Titanium (94% energy efficiency)
- Cooling: Custom (soft piped) Water Cooling for both the CPU and GPUs: a refilling hole drilled in the top of the chassis, and transparent reservoir in the front (see pictures below)

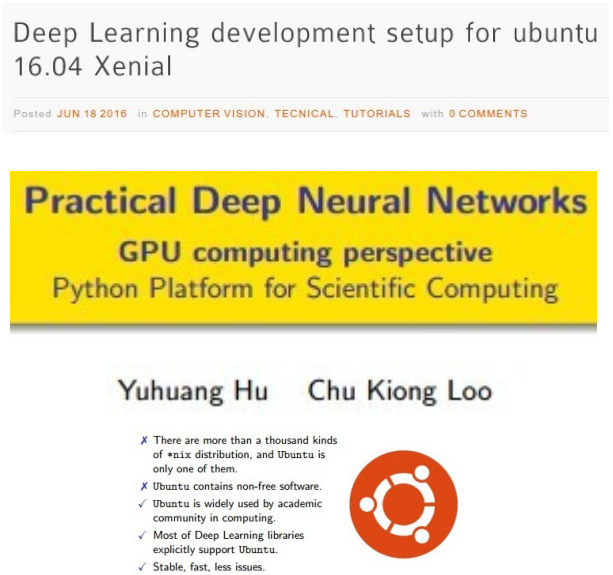
a beautiful sight... left: The system is being built. You can see the plastic piping for the water cooling going through the holes already available in the Carbide Air 540 chassis. The motherboard is vertically mounted.
middle & right: The system is completely built. Notice that the water reservoir can be seen from the outside. Red plastic pipes can be seen going from up (there is a filling hole on the outside), down to the water pump, through the water blocks installed on the GPUs (keeping these cool). A similar thing happens for the CPU which has its separate cool block and pipes leading to and from it.

http://graphific.github.io/posts/building-a-deep-learning-dream-machine/

Operating System UBUNTU 14.04 OR 16.04

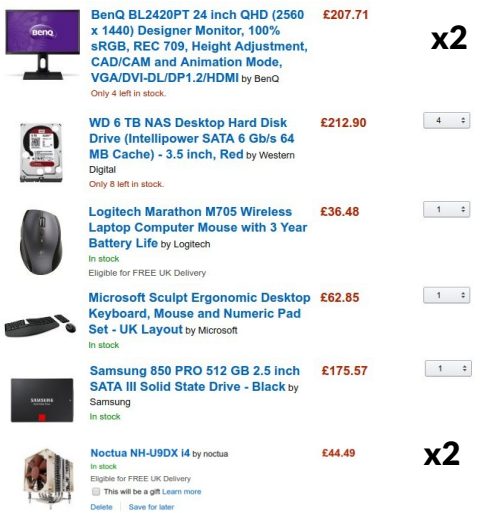
- CUDA Toolkit 8.0 has native support for versions 14.04 and 16.04
  - 14.04 LTS might be still a safe choice
- NVIDIA CUDA is not always the most straightforward to make work in Ubuntu
  - Easy to break your graphic drivers, so have an Ubuntu Live CD available.
    - Prepare for Glib -related issues as well

CUDA Toolkit 8.0



Base Build COMMON COMPONENTS

- Basically the main differences between the 4 different options are in the choice of:
  - Motherboard
  - CPU
  - GPU
  - Case+Power
- The following shopping list can be defined initially as base “investment” for our system (of course these should be tailored for your needs) that should work with all the chosen options. See details later



Dual-monitor setup with basic 2560 x 1440 IPS monitor

18 TB with RAID 5, if you need less, you save money

512 GB for the Operating system

~ £1,600

“CONSUMER” Shopping List INTEL XEON DUAL-CPU

- Ballpark estimate of the “typical” consumer build for deep learning.

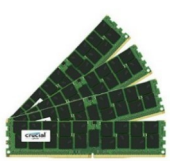
Intel Xeon E5-2620 v4 S 2011-3 Broadwell-EP 8 Core 20 MB Processor by Intel — £458.11 — 2

Usually dispatched within 2 to 3 days
Sold by LambdaTek ComponentShop
Gift options not available. Learn more
Delete | Save for later

ASUS Z10PE-D16 WS - motherboard - SSI EEB - LGA2011-v3 Socket - C612 by ASUS — £483.03

Only 5 left in stock.
Is delivered in Certified Frustration-Free Packaging
Eligible for FREE UK Delivery

6 x PCIe slots only
→ 3 GPUs only



Crucial CT4K16G4RFD424A DDR4, 64 GB (4 x 16 GB), DIMM, 288-Pin, 2400 MHz, PC4-19200, CL17, 1.2 V Internal Memory
by Crucial
Be the first to review this item
RRP: £762.00
Price: £320.50 FREE UK delivery.
You Save: £441.50 (58%)
Only 2 left in stock - order soon.
Estimated delivery 16 - 18 Aug. when you choose Express Delivery at checkout.

2,133 MHz PC4-17000 memory is £20 cheaper per 64 GB

NOTE!
Pascal Titan X released on Aug 2, 2016
$1200 (**~£920**). **Limited to 2 per customer!**

Fractal Design Define FD-CA-DEF-XL-R2-BL PC Casing ATX / 4 x 5.25 / 6 x 3.5 / 2 x USB 3.0 / Pearl Black by Fractal Design — £105.45 — 1

Only 12 left in stock.

EVGA SuperNOVA 1600W Modular Power Supply
120-G2-1600-X3 - 1600W EVGA SUPERNOVA 1600 G2, Full Modular, 80 PLUS Gold, 140mm Quiet Fan, ATX, PSU
£249.98 Inc VAT **NOT FOUND FROM AMAZON**

Samsung 950 PRO NVMe M.2 512 GB Solid State Drive - Black by Samsung — £275.99 — 1

Only 8 left in stock (more on the way).
Eligible for FREE UK Delivery

Faster “cache” M.2 SSD
or simply faster OS disk

Intel Xeon BASE Price
~ £1600 previous slide

+£920	Xeon-specific
+£480	~£2,600
+£1,280	
+(3x£920)	Xeon and i7
+£100	~£3,400
+£250	
+£275	

~ £7,700

“CONSUMER” Shopping List DUAL-CPU → SINGLE-CPU SYSTEM

XEON dual CPU

ASUS Z10PE-D16 WS - motherboard - SSI EEB - LGA2011-v3 Socket - C612 by ASUS — £483.03 — 1

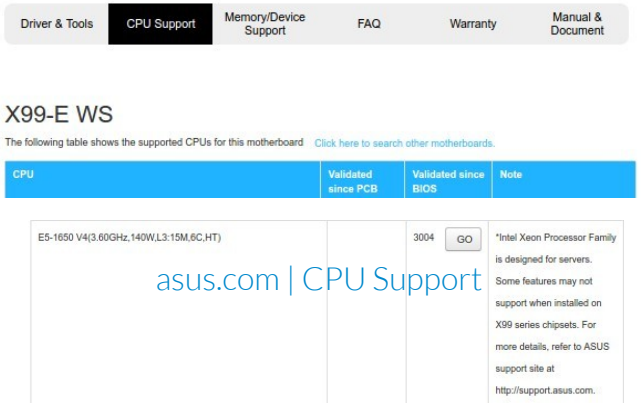
Only 5 left in stock.
Is delivered in Certified Frustration-Free Packaging
Eligible for FREE UK Delivery

Intel Xeon E5-2620 v4 S 2011-3 Broadwell-EP 8 Core 20 MB Processor by Intel — £458.11 — 2

Usually dispatched within 2 to 3 days
Sold by LambdaTek ComponentShop
Gift options not available. Learn more
Delete | Save for later

HyperX FURY DDR4 HX421C14FBK4/64 Ram Kit 64 GB (4 x 16 GB) 2133 MHz DDR4 CL14 DIMM — £267.10
by HyperX
Only 1 left in stock (more on the way).

~£40 difference to EEC RAM (per 64 GB) for i7 setup



asus.com | CPU Support

Single i7 CPU

HyperX FURY DDR4 HX421C14FBK4/64 Ram Kit 64 GB (4 x 16 GB) 2133 MHz DDR4 CL14 DIMM — £267.10
by HyperX
Only 1 left in stock (more on the way).

Intel i7 5930K CPU Processor (3.50GHz, 15MB Cache, 140W, Socket 2011-V3) by Intel — £504.20
Only 2 left in stock.
Sold by Boss-Computer GBR
For Peteri Teikari's Wish List
Gift options not available. Learn more
Delete | Save for later

Asus X99-E WS Workstation Motherboard (Intel X99, DDR4, S-ATA 600, CEB, 1 x M.2, Quad Strength Graphic Power, 12K Hours Capacitors, Socket 2011-v3) by ASUS — £399.99
Only 1 left in stock (more on the way).
Eligible for FREE UK Delivery
For Peteri Teikari's Wish List
This will be a gift. Learn more
Delete | Save for later

8 x PCIe slots
→ 4 GPUs

~£500 cheaper
with single i7 CPU and the Asus X99-E Motherboard

~ £1,600 + £3,400 + (£400+£500) + £920 **~£6,800**
base “consumer” CPU+MB Titan X

Single Xeon CPU, E5-16xx

Asus X99-E WS Workstation Motherboard (Intel X99, DDR4, S-ATA 600, CEB, 1 x M.2, Quad Strength Graphic Power, 12K Hours Capacitors, Socket 2011-v3)
Pay Monthly available over £399.99
by ASUS
6 customer reviews | 7 answered questions
Price: £399.99 & FREE Delivery in the UK. Details
Only 2 left in stock (more on the way).

8 x PCIe slots
→ 4 GPUs

~£425 cheaper
with single E5-16xx CPU and the Asus X99-E Motherboard

~ £7,700-£425+ £920 **~£8,200**

Shopping List 2ND HAND GPUS/PC

- If you are tight on money, you could consider “renting” 2nd hand GPUs by buying them from eBay and selling them away after some while instead of using extensively cloud services. In the long-term this is a lot more economical if you can manage the upfront cost.

reddit.com/r/MachineLearning:

r4and0muser9482 1 point 4 days ago

If yuou buy used and use it, it's practically worthless when you want to sell it. You either buy new and sell used or buy used and throw into the bin. On the other hand, most people buy new and use until it's worthless. You don't change graphic cards every 6 months.

sentdex 2 points 4 days ago

Why is it practically useless when you want to sell it? I've been buying processors and GPUs for years, running them nearly 24/7, and then selling after about a year.

Haven't experienced this "useless" thing you're talking about. Maybe you don't change graphics cards often, but you can, and it's far cheaper than buying cloud computing.

r4and0muser9482 1 point 4 days ago

Well, I may be unexperienced. Why don't you give some examples?

I do buy quite a bit of hardware, but I'm more of a buy-new-throw-away kind of person.

[–]sentdex 1 point (6 August 2016)

Sure, gtx 980 was ~ 400ish USD used a year ago, ~300ish used today. Titan X maxwell is currently $700-$750 used today, I expect it to be worth ~400-500 in a year from now. Certainly not "worthless" as you put it. K80 at 0.56 per hour? That's $13 USD day, and $4,905.60 in a year. That's pure loss, that money is gone.

If you go buy a new K80, that's $4,000 USD, but that's TWO K80 GPUs in one, where it seems like Azure is splitting up the card, so really you're just getting half that, or a $2,000 value. Even if you got the full $4,000 value, in a year, you've paid an extra $1,000, even if the value of the card is worthless as you put it by the end of the year.

It wont be worthless after a year, however, the card will likely be more like $2000 USD used. So, after a year, you can either spend $4,905 (or, what looks more like double that, since Azure is selling half the K80 at a time since it's 2 GPUs, so more like $9,810), or you can spend $2,000 USD + electricity. That's if you buy new.

Buy used, and then we're compared the likely $9,810 to say a card you buy for $2,000 and then sell a year later for $1500, losing $500. You can also buy K80s right now used for 3K usd. I could be wrong on the calc, I didn't see prices, but in their table they claim a K80 gpu as being really the half of a K80 gpu since it's a dual GPU card. These are major differences here. While Azure is better than AWS, it's still massively more expensive. If cost is your main concern, buy used, sell used.

If you only train models once in a while, then cloud might actually be a great place to be, rather than buying a card. If you are continually training though, just doesn't make much sense.

Shopping List CUTTING COSTS?

- For further savings, you can look for new or 2nd hand AMD Opteron CPUs that are designed as competitors for Intel Xeon range

May 7, 2015
AMD Refreshes Roadmap, Transitions Back to HPC
Tiffany Trader

hpcwire.com/2015/05/07

Advantages of Opterons

http://www.tomshardware.co.uk/forum/343718-28-xeon-opteron-comparison

- Significantly **less expensive** than anywhere near similar Xeons, especially the quad-CPU-capable Opteron 6200s vs. the quad-CPU-capable Xeon E5-4xxx or E7-4xxx CPUs.
- Opterons are sold by core count and clock speed, everything else (feature set, cache size, RAM speed, bus speed) is identical between processors in the same model line. There are no "crippled" Opterons. Not so with Xeons, only the really expensive ones are "fully functional" with full-speed buses, full-speed memory controllers, all of the whiz-bang features enabled, the full amount of cache, etc.
- Performs much better than a similarly priced Xeon CPU especially using Linux or any other non-Windows OS, and especially on multithreaded code.
- AMD tends to support a CPU socket for more than 1-2 generations, so if you buy an AMD server board, you are more likely to be able to upgrade the CPU for a longer period of time without shelling out for a new board.
- Boards tend to be a little less expensive than equivalent Intel boards
- Greater supported memory bandwidth- all Opterons support up to DDR3-1866, while no Xeon can support more than DDR3-1600, and many multi-CPU ones only support DDR3-1333.
- Lower power consumption with quad-socket CPUs, especially the Opteron 6200s vs. Xeon E7s due to the latter's FB-DIMM2 on-motherboard memory buffers.

Advantages of Xeons

- Somewhat greater selection of motherboards available
- Better single-threaded performance than Opterons, especially with legacy Windows applications
- Can scale up to 8 CPU systems (E7-8xxx) whereas current Opterons only support up to 4 CPU systems.
- 8 CPU Xeon E7s can support more RAM per board than 4 CPU Opteron setups.
- Nobody ever got in trouble for buying Intel. Buying from the 800 lb gorilla of the market allows the IT peon to shift the blame from themselves if anything goes wrong. Buying from somebody other than the 800 lb gorilla in the field invites personal criticism and blame if anything goes wrong even if saves the CFO a bundle.

PASSMARK | Multiple CPU Systems

[Dual CPU] AMD Opteron 6276, 10947 CPU Mark, £140

[Quad CPU] AMD Opteron 6276, 15342 CPU Mark, £280

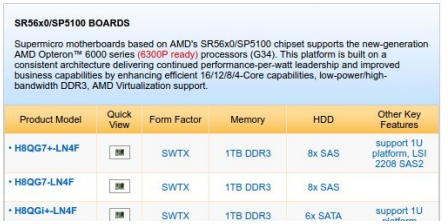
[Dual CPU] Intel Xeon E5-2620 v4 , 17655 CPU Mark, £920

Rather good performance for the price

See "Quad Xeon vs Opteron, Zemax OpticStudio" by Dr Donald Kinghorn of Puget Systems

- For NVIDIA GPUs, you don't have in practice an alternative from AMD, very limited use for deep learning with Radeon GPUs and OpenCL.
  - amd/OpenCL-caffe
  - OpenCL / AMD: Deep Learning
  - Deep Learning on OpenCL
  - reddit.com/r/MachineLearning | OpenCL and AMD

Opteron™ 6000 Series (6300P ready) Processor-based Motherboards
supermicro.com/Aplus/motherboard/Opteron6000/



SR56x0/SP5100 BOARDS

Supermicro motherboards based on AMD's SR56x0/SP5100 chipset supports the new-generation AMD Opteron™ 6000 series (6300P ready) processors (G34). This platform is built on a consistent architecture delivering continued performance-per-watt leadership and improved business capabilities by enhancing efficient 16/12/8/4-Core capabilities, low-power/high-bandwidth DDR3, AMD Virtualization support.

Product Model	Quick View	Form Factor	Memory	HDD	Other Key Features
H8QG7+-LN4F		SWTX	1TB DDR3	8x SAS	support 1U platform, LSI 2208 SAS2
H8QG7-LN4F		SWTX	1TB DDR3	8x SAS	
H8QGi+-LN4F		SWTX	1TB DDR3	6x SATA	support 1U

GPU: Memory vs. Performance

COMPARED TO MAXWELL TITAN X

NVIDIA · Graphics Cards · Deep Learning

What are the downsides of the NVIDIA's GTX 1080 vs Titan X when used for deep learning?

quora.com

Djébril Mokaddem, Deep Learning R&D Engineer
2.9k Views

Be careful guys. The 1080 has indeed the 16-bit support but these instructions are 1/64th slower than the 32-bits's. IMO, this GPU isn't a high-end Pascal chip (GP104), but only transitional to the ultimate GP100 that the Tesla P100 uses.

So, if you think you will get 16 gigs on the 1080 versus 12 on the titan at the same or greater speed, think again. You should wait for the next titan, whose according to chinese leaks, will be based on the GP100/GP102 with ... buckles up ... 24 gigs !

In conclusion, if you have to buy a card now, you should go for the titan x.

Written May 29 • View Upvotes

reddit m.reddit.com/r/buildapc

r/buildapc • Build Help

[Build Help] Deep Learning PC with GTX 1080

u/zubz12 • 33d, 14h

dufu · 29d, 8h

I just built a rig for the GTX 1080 for use with Tensorflow. Your CPU and board are probably overkill. I know the 16 pcie lanes limit is a concern but I'm pretty sure you can run 2 gpus in 8x without much if any lag. Honestly, I bet one GTX 1080 would be sufficient for your needs. Here's my build for ~$700 (excluding GPUs). Only 168 watts too and doesn't require additional cooling. PCPartPicker part list / Price breakdown by merchant

GEAR & GADGETS / PRODUCT NEWS & REVIEWS

Nvidia GTX 1080 review: Faster, cheaper, quieter than Titan X

Already on top, Nvidia pushes the performance curve once again—but has it gone far enough?

by **Mark Walton** - May 17, 2016 1:59pm BST

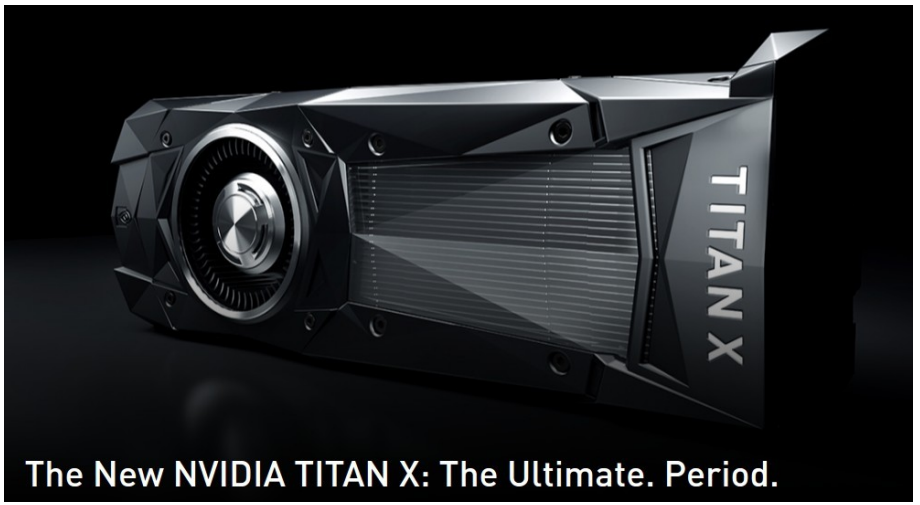
Share Tweet Email 298

http://arstechnica.co.uk/gadgets/2016/05/nvidia-gtx-1080-review/

GPU2: Memory and Performance

AFTER PASCAL TITAN X – BEST VALUE OF MONEY

- With the introduction of the Pascal-based Titan X (note the confusing naming), you get 12 GB of memory and outperforming GTX1080 at the same time.
- So **buy definitely Titan X** (Pascal, v. 2016) unless you want to wait for the GTX 1080 Ti for more options (pushes down the prices of GTX1060/1070/1080 maybe?)



https://blogs.nvidia.com/blog/2016/07/21/titan-x/
https://www.reddit.com/r/nvidia/comments/4u0mbv/the_new_nvidia_titan_x_the_ultimate_period
https://news.ycombinator.com/item?id=12141334

NVIDIA GeForce 10 Pascal Family:

Graphics Card Name	NVIDIA GeForce GTX 1060 6 GB	NVIDIA GeForce GTX 1070	NVIDIA GeForce GTX 1080	NVIDIA GeForce GTX Titan X
Graphics Core	GP106	GP104	GP104	GP102
Process Node	16nm FinFET	16nm FinFET	16nm FinFET	16nm FinFET
Die Size	200mm2	314mm2	314mm2	TBD
Transistors	4.4 Billion	7.2 Billion	7.2 Billion	12.0 Billion
CUDA Cores	1280 CUDA Cores	1920 CUDA Cores	2560 CUDA Cores	3584 CUDA Cores
Base Clock	1506 MHz	1506 MHz	1607 MHz	TBD
Boost Clock	1708 MHz	1683 MHz	1733 MHz	1530 MHz
FP32 Compute	4.4 TFLOPs	6.5 TFLOPs	9.0 TFLOPs	11 TFLOPs
VRAM	6 GB GDDR5	8 GB GDDR5	8 GB GDDR5X	12 GB GDDR5X
Bus Interface	192-bit bus	256-bit bus	256-bit bus	384-bit bus
Power Connector	Single 6-Pin Power	Single 8-Pin Power	Single 8-Pin Power	8+6 Pin Power
TDP	120W	150W	180W	250W
Display Outputs	3x Display Port 1.4 1x HDMI 2.0b 1x DVI	3x Display Port 1.4 1x HDMI 2.0b 1x DVI	3x Display Port 1.4 1x HDMI 2.0b 1x DVI	3x Display Port 1.4 1x HDMI 2.0b 1x DVI
Launch Date	13th July 2016	10th June 2016	27th May 2016	2nd August 2016
Launch Price	$249 US	$379 US	$599 US	$1200 US

Note! The price difference though, and the increased power demand in comparison to GTX1080 (same TDP of 250W for the previous generation Titan X though)

GPU3: More Memory

NVIDIA Quadro Specification Comparison				
	P6000	P5000	M6000	M5000
CUDA Cores	3840	2560	3072	2048
Texture Units	240?	160	192	128
ROPs	96?	64	96	64
Core Clock	?	?	N/A	N/A
Boost Clock	~1560MHz	~1730MHz	~1140MHz	~1050MHz
Memory Clock	9Gbps GDDR5X	9Gbps GDDR5X	6.6Gbps GDDR5	6.6Gbps GDDR5
Memory Bus Width	384-bit	256-bit	384-bit	258-bit
VRAM	24GB	16GB	24GB	8GB
FP64	1/32 FP32	1/32 FP32	1/32 FP32	1/32 FP32
TDP	250W	180W	250W	150W
GPU	GP102	GP104	GM200	GM204
Architecture	Pascal	Pascal	Maxwell 2	Maxwell 2
Manufacturing Process	TSMC 16nm	TSMC 16nm	TSMC 28nm	TSMC 28nm
Launch Date	October 2016	October 2016	03/22/2016	08/11/2015
Launch Price (MSRP)	TBD	TBD	$5000	$2000

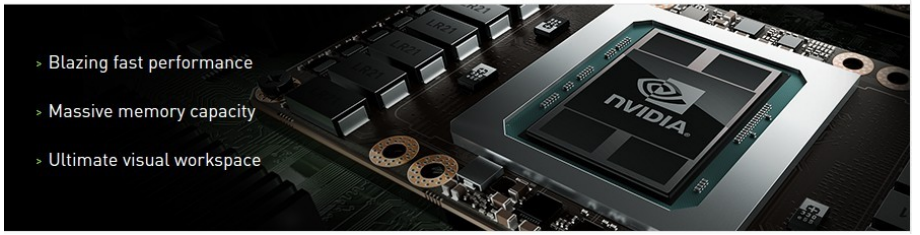
$5000?

QUADRO P6000 SPECS	
CUDA Parallel-Processing Cores	3840
GPU Memory	24 GB GDDR5X
FP 32 Performance	12 TFLOPs
Max Power Consumption	250 W
Graphics Bus	PCI Express 3.0 x 16
Display Connectors	DP 1.4 (4) DVI-D (1) Optional Stereo (1)
Form Factor	4.4" H x 10.5" L Dual Slot

QUADRO P5000 SPECS	
CUDA Parallel-Processing Cores	2560
GPU Memory	16 GB GDDR5X
FP 32 Performance	8.9 TFLOPs
Max Power Consumption	180 W
Graphics Bus	PCI Express 3.0 x 16
Display Connectors	DP 1.4 (4) DVI-D (1) Optional Stereo (1)
Form Factor	4.4" H x 10.5" L Dual Slot

Designs are becoming more complex. Media is becoming richer with higher fidelity, combining greater resolutions and complex visual effects. Scientific visualization and compute problems are larger than ever. Virtual Reality (VR) is changing all facets of entertainment, design, engineering, architecture, and medicine. Professionals want to experience ideas, validate designs, rehearse procedures, and visualize problems interacting with them naturally and at scale.

Fueled by NVIDIA Pascal™, NVIDIA's most powerful GPU architecture ever, the new Quadro products bring a whole new level of performance and innovative capabilities to power visual computing on the desktop, in VR, or on-the-go. Whether you're creating revolutionary products, designing ground-breaking architecture, or telling spectacularly vivid visual stories, Quadro gives you the power to do it better and faster.



http://www.nvidia.com/object/quadro-graphics-with-pascal.html

IMPRESSIVE AMOUNT OF MEMORY

If you have 8 x $5,000 to invest, get this with server MB, and you will get 192 GB of memory with 96 TFLOPS (FP32) for medical volumetric images (MRI, optical and electron microscopy, OCT, etc.)

GPU: What about the Teslas?

NVIDIA Tegra | NVIDIA GeForce | NVIDIA | Deep Learning | +3

Would you build a multi-GPU system for deep learning with GTX Titan X or Tesla K40/K80? What are the pros and cons?

Update Dec 23, 2015: I ended up going with multiple GTX Titan X in a single box. I'm really happy with it so far. Tesla K40/80 & InfiniBand together are just too expensive. As I'm running this box in my basement, those server features aren't a big deal for me. GTX Titan X memory & # cores are great!

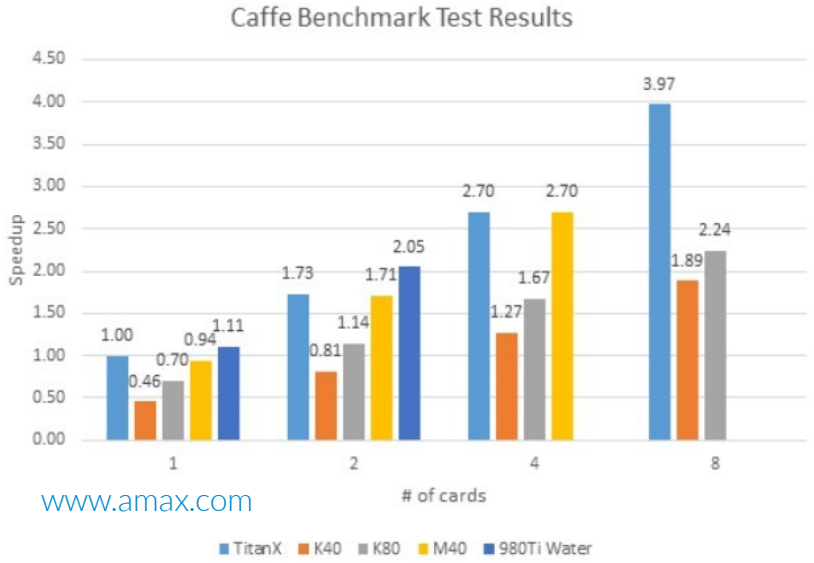
https://www.quora.com/

You find Teslas used in many older deep learning papers, but they are designed mainly for high precision (double, FP64) scientific computations,

whereas for **deep learning** purposes **single precision** is sufficient (FP32)

Basic Performance Analysis of NVIDIA GPU Accelerator Cards for Deep Learning Applications

Frank Han, Thomas Zhu, and Rene Meyer



www.amax.com

Microway
Technology you can count on since 1982.
HOM

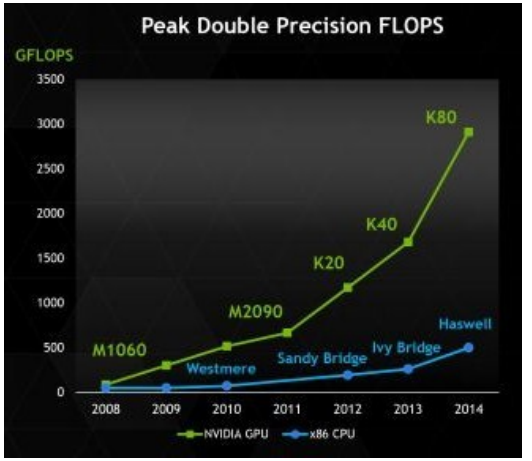
Why Microway | Markets Served | Technologies | Products | Knowledge Center

HOME > KNOWLEDGE CENTER > RESOURCES > COMPARISON OF NVIDIA TESLA AND NVIDIA GEFORCE GPUS

Comparison of NVIDIA Tesla and NVIDIA GeForce GPUs

This resource was prepared by Microway from data provided by NVIDIA.

microway.com



http://www.anandtech.com

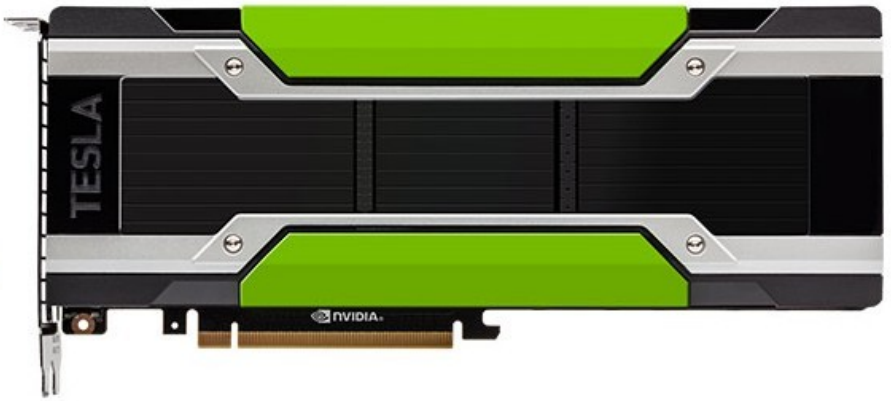
GPU: 16-bit Floating Point

With GTX1080 / Titan X / Quadro P6000 you can compute with **half-precision floats (FP16)** but it is actually slower than with single-precision (FP32).

NOT THE CASE WITH P100

NVIDIA TESLA P100 FOR MIXED-WORKLOAD HPC

Tesla P100 for PCIe enables mixed-workload HPC data centers to realize a dramatic jump in throughput while saving money. For example, a single GPU-accelerated node powered by four Tesla P100s interconnected with PCIe replaces up to 32 commodity CPU nodes for a variety of applications. Completing all the jobs with far fewer powerful nodes means that customers can save up to 70% in overall data center costs.



PERFORMANCE SPECIFICATION FOR NVIDIA TESLA P100 ACCELERATORS

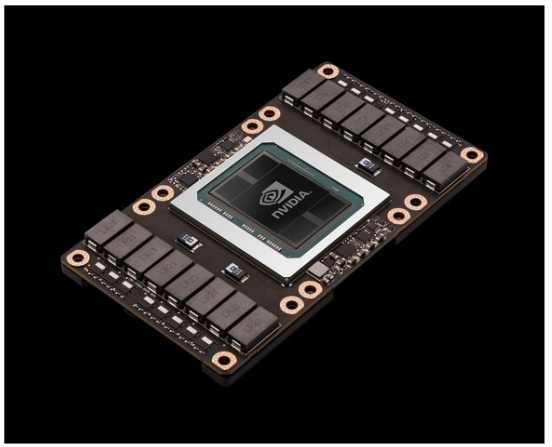
	P100 for PCIe-Based Servers	P100 for NVLink-Optimized Servers
Double-Precision Performance	4.7 TeraFLOPS	5.3 TeraFLOPS
Single-Precision Performance	9.3 TeraFLOPS	10.6 TeraFLOPS
Half-Precision Performance	18.7 TeraFLOPS	21.2 TeraFLOPS
NVIDIA NVLink™ Interconnect Bandwidth	-	160 GB/s
PCIe x16 Interconnect Bandwidth	32 GB/s	32 GB/s
CoWoS HBM2 Stacked Memory Capacity	16 GB or 12 GB	16 GB
CoWoS HBM2 Stacked Memory Bandwidth	720 GB/s or 540 GB/s	720 GB/s
Enhanced Programmability with Page Migration Engine	✓	✓
ECC Protection for Reliability	✓	✓
Server-Optimized for Data Center Deployment	✓	✓

http://www.nvidia.com/object/tesla-p100.html

Home / Hardware

Servers with Nvidia's Tesla P100 GPU will ship next year

The new GPU will provide a serious performance boost to servers and supercomputers



Nvidia's Tesla P100 graphics processor is based on the Pascal architecture.
Credit: Nvidia

Agam Shah Apr 5, 2016 1:31 PM
IDG News Service

PCWorld FROM IDG
pcworld.com

Nvidia's GPUs are widely used in supercomputers today. Two of the world's 10 fastest supercomputers use Nvidia GPUs, according to a list compiled by Top500.org.

DGX-1 8x P100 GPUs

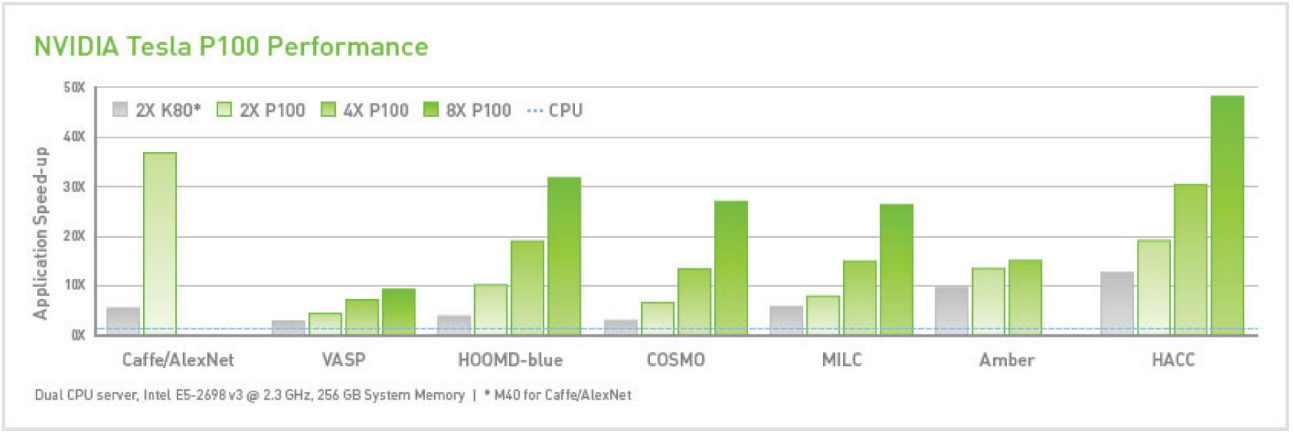
THE WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER IN A BOX

ORDER YOUR NVIDIA DGX-1

The NVIDIA DGX-1 is available for purchase in select countries and is priced at $129,000*. DGX-1 service and support at additional cost.

NVIDIA DGX-1, http://www.nvidia.com/object/deep-learning-system.html

Jen-Hsun Huang and Elon Musk over DGX-1, "Delivering World's First AI Supercomputer in a Box to OpenAI" - blogs.nvidia.com



http://www.nvidia.com/object/tesla-p100.html

SYSTEM SPECIFICATIONS	
GPUs	8x Tesla P100
TFLOPS (GPU FP16 / CPU FP32)	170/3
GPU Memory	16 GB per GPU
CPU	Dual 20-core Intel® Xeon® E5-2698 v4 2.2 GHz
NVIDIA CUDA® Cores	28672
System Memory	512 GB 2133 MHz DDR4
Storage	4x 1.92 TB SSD RAID 0
Network	Dual 10 GbE, 4 IB EDR
Software	Ubuntu Server Linux OS DGX-1 Recommended GPU Driver
System Weight	134 lbs
System Dimensions	866 D x 444 W x 131 H (mm)
Packing Dimensions	1180 D x 730 W x 284 H (mm)
Maximum Power Requirements	3200W
Operating Temperature Range	10 - 30° C

170 TFLOPS (8*21.2, FP16)
84.8 TFLOPS 8*10.6, FP32)

VS.

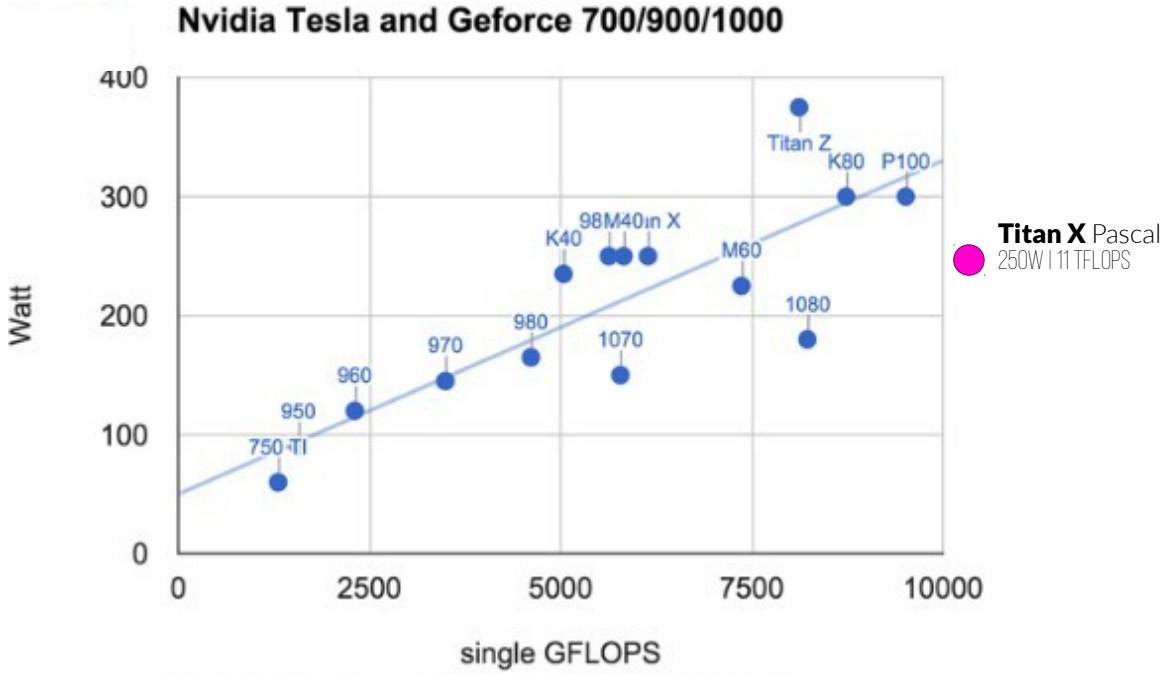
11 FLOPS (FP32) OF **TITAN X**

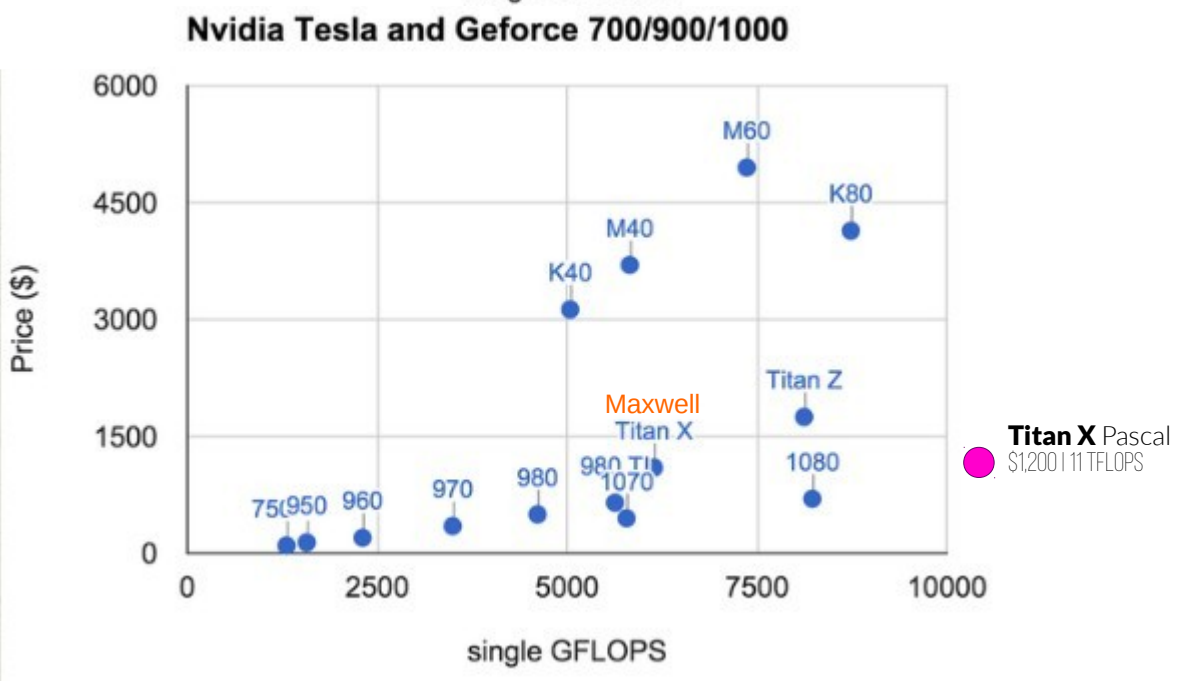
GPU PERFORMANCE/PRICE

Modified graphs from Guanghan Ning's report that did not include the recently released **Titan X Pascal** that now dominates both of the graphs.

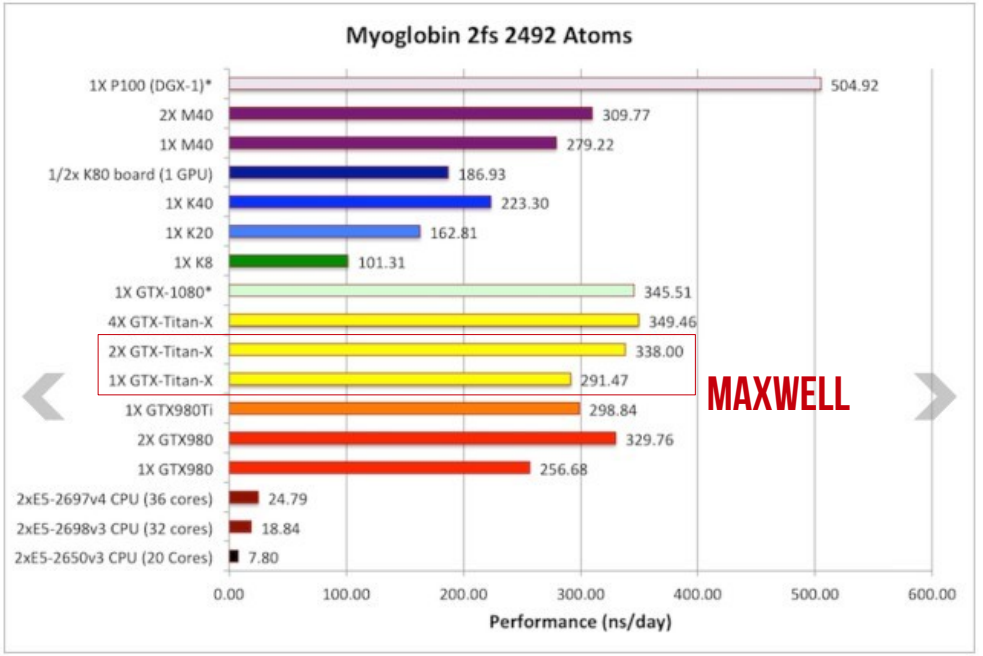
NVIDIA **Quadro P6000** achieves 12 TFLOPS at same 250 W with a ~$5,000 price tag (not sure yet).

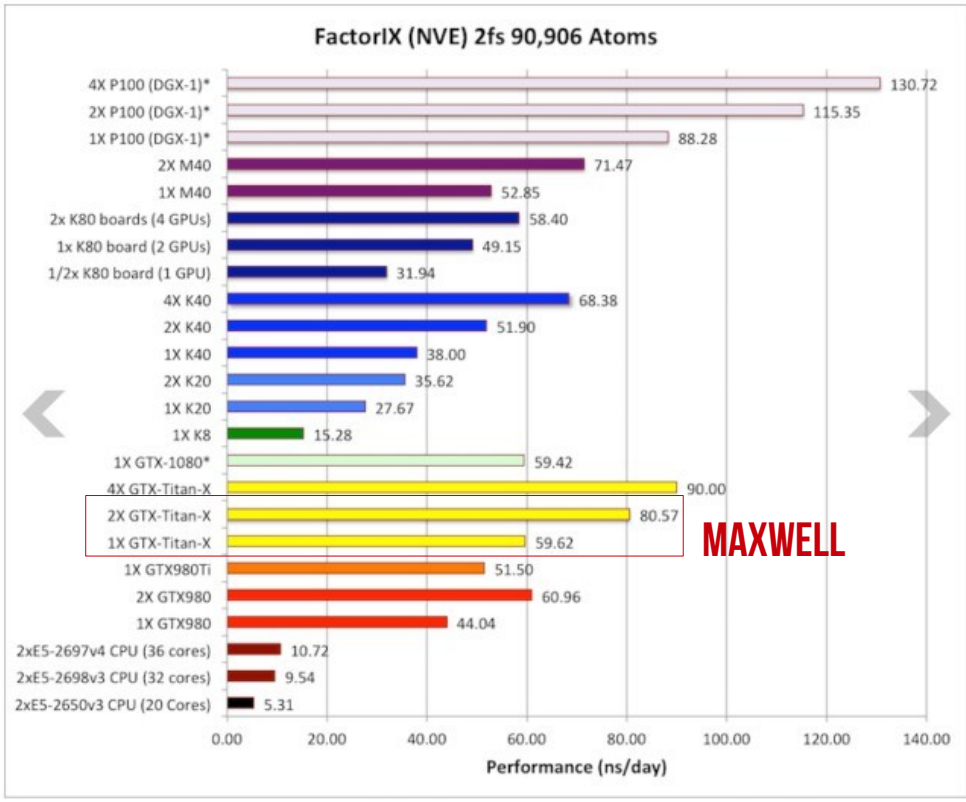
The **Tesla** line (K40, M40, M60, K80) just give a very poor performance in deep learning compared to the price.



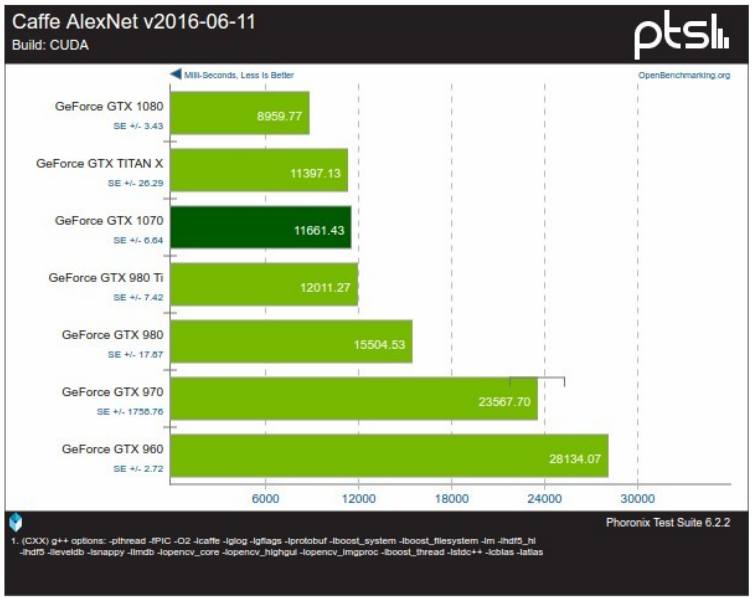


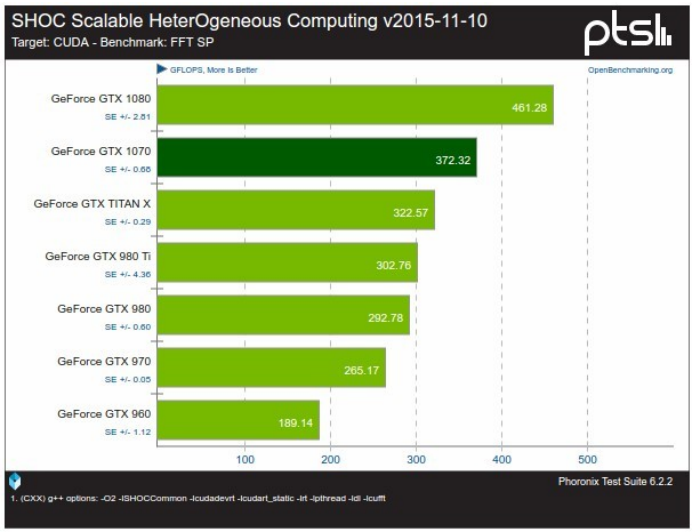
GPU Benchmarks





Non-deep learning scientific calculations, http://wccftech.com/nvidia-gp100-gpu-tesla-p100-benchmarks/



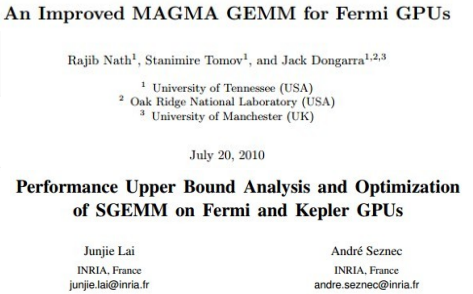


phoronix.com

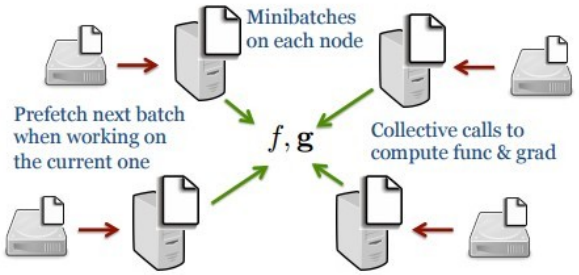
Batch size PRACTICAL BENCHMARKING

- Deep learning trained typically using minibatches. Bigger minibatches more efficient to compute, but require more GPU memory as well.
  - In practice the computation optimization is based on a function called GEMM (GEneral Matrix to Matrix Multiplication). It's part of the BLAS (Basic Linear Algebra Subprograms) library that was first created in 1979. Pete Warden: "Why GEMM is at the heart of deep learning"

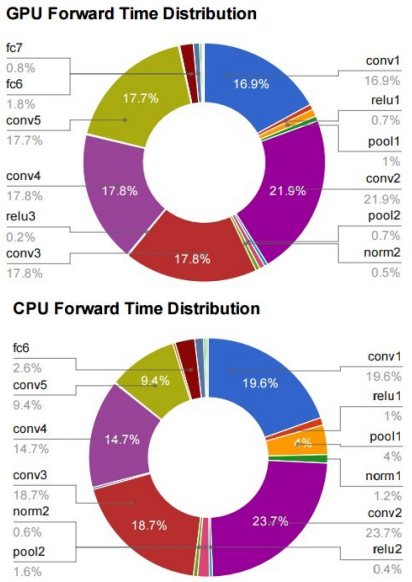
YISONG YUE'S BLOG: Minibatches: Use minibatches. Modern computers cannot be efficient if you process one training case at a time. It is vastly **more efficient to train the network on minibatches** of 128 examples, because doing so will result in massively greater throughput. It would actually be nice to use minibatches of size 1, and they would probably result in improved performance and **lower overfitting**; but the benefit of doing so is outweighed the massive computational gains provided by minibatches. But don't use very large minibatches because they tend to work less well and **overfit more**. So the practical recommendation is: use the smaller minibatch that runs efficiently on your machine.



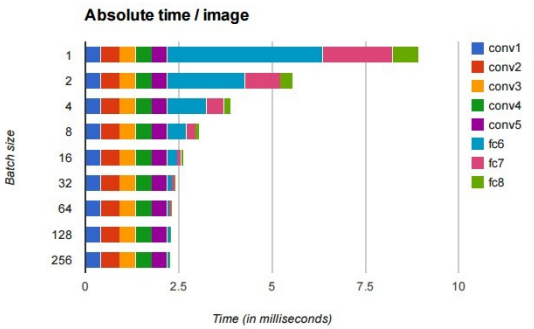
Jia (2014): **"Learning Semantic Image Representations at a Large Scale"**



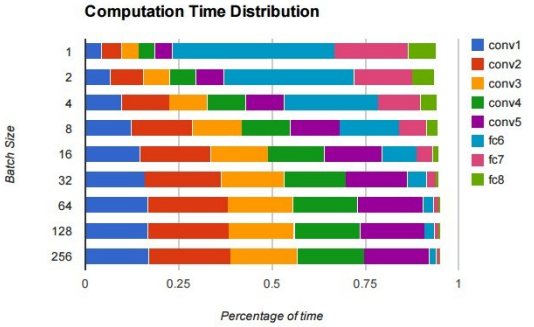
We took advantage of parallel computing by distributing the data over multiple machines and performing gradient computation in parallel, as it only involves summing up the perdatum gradient. As the data is too large to fit into the memory of even a medium-sized cluster, we only keep the minibatch in memory at each iteration, with a background process that **pre-fetches the next minibatch** from disk during the computation of the current minibatch. This enables us to perform efficient optimization with an arbitrarily large dataset.



Computation time distribution of individual layers, on both GPUs and CPUs for the forward pass.



Top: absolute computation time spent on each layer per image, when the batch size varies from 1 to 256.



Bottom: an alternative view of the top chart, showing the relative percentage of computation time by each layer. Note that in both figures, low-cost layers such as pooling and ReLU are not shown, which accounts for the 5% unoccupied percentage in the bottom figure.

Batch size PRACTICAL BENCHMARKING #2

- So what does this mean in terms of GPU selection?
  - Minibatch should fit GPU memory. See table on right. Doubling the batch size (128 → 256) roughly halves the computation time. But if that does not fit into memory you can't get the speedup.

GPUs	Batch size	Cross-entropy	Top-1 error	Time	Speedup
1	(128, 128)	2.611	42.33%	98.05h	1x
2	(256, 256)	2.624	42.63%	50.24h	1.95x
2	(256, 128)	2.614	42.27%	50.90h	1.93x
4	(512, 512)	2.637	42.59%	26.20h	3.74x
4	(512, 128)	2.625	42.44%	26.78h	3.66x
8	(1024, 1024)	2.678	43.28%	15.68h	6.25x
8	(1024, 128)	2.651	42.86%	15.91h	6.16x

Batch size (*m*, *n*) indicates an effective batch size of *m* in the convolutional layers and *n* in the fully-connected layers. Krizhevsky (2014)

- Go for GPUs with bigger memory as in practice you will never have too much memory on your GPU

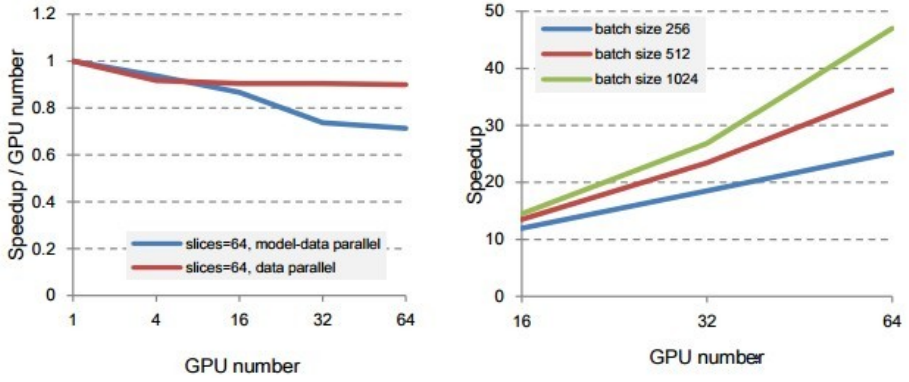


Figure 3: **Left:** The scalability of different parallel approaches. The hybrid parallelism is better when number of GPUs is less than 16. The scalability of data parallelism is better with large numbers of GPU because the communication consuming is constant. **Right:** The speedup of going through images. The larger the batch size is, the larger the speedup is.

Wu et al. (2015)

We tested the scaling efficiency by training a model for an image classification task. The network has 8 convolutional layers and 3 fully-connected layers followed by a 1000-way softmax. Measured by the epoch time, the scalability efficiency and the speedup of going through images are shown in Figure 3. For the convenience of observing the scalability of **different parallel strategies**, we fixed the number of images processed by each GPU to 64 (slices = 64).

The time taken for **hybrid parallelism** and **data parallelism** with different numbers of GPUs is shown in the left-hand figure. The data parallelism performs better when the involved GPU number is larger than 16. This is because communication overhead of the data parallel strategy is constant when the size of model is fixed. The speedup is larger with larger batch size as shown in the right-hand figure. Compared with a single GPU, a 47x speedup of going through images is achieved by using 64 GPUs with a mini-batch size of 1024. As the number of GPUs increases, the total device memory is also increasing, and more data can be cached on the device memory. This is helpful for improving parallel efficiency.

GPU Benchmarks #2 WITH GPU MEMORY

Benchmarks for popular CNN models

https://github.com/jcjohnson/cnn-benchmarks by Justin Johnson

ResNet-50

(input 16 x 3 x 224 x 224)

This is the 50-layer model described in [3] and implemented in fb.resnet.torch.

GPU	Forward (ms)	Backward (ms)	Total (ms)
TITAN X (cuDNN 5005)	36.25	73.93	110.18
GeForce GTX 1080 (cuDNN 5005)	50.67	103.24	153.90
GeForce GTX TITAN X (cuDNN 5005)	56.42	114.60	171.02
TITAN X (nn)	87.81	161.03	248.83
GeForce GTX 1080 (nn)	109.81	201.66	311.47
GeForce GTX TITAN X (nn)	136.37	245.99	382.36
CPU: Dual Intel Xeon E5-2630 v3	2477.61	4149.64	6627.25

Maxwell Titan X (Geforce Titan X) is slower than GTX1080

ResNet-200

(input 16 x 3 x 224 x 224)

This is the 200-layer model described in [4] and implemented in fb.resnet.torch.

Even with a batch size of 16, the 8GB GTX 1080 did not have enough memory to run the model.

GPU	Forward (ms)	Backward (ms)	Total (ms)
TITAN X (cuDNN 5005)	104.98	205.82	310.80
GeForce GTX TITAN X (cuDNN 5005)	171.15	322.66	493.82
TITAN X (nn)	313.90	522.16	836.05
GeForce GTX TITAN X (nn)	491.69	806.95	1298.65
CPU: Dual Intel Xeon E5-2630 v3	8666.43	13758.73	22425.16

Maxwell Titan X (Geforce Titan X) here can at least be used with the batch size of 16.

With GTX 1080 you could not train this deep model, as it does not have enough memory (8 GB vs. 12 GB of Titan X)

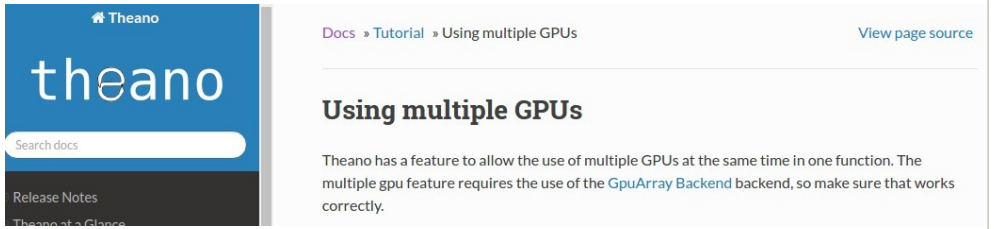
Some general conclusions from this benchmarking:

- Pascal Titan X > GTX 1080: Across all models, the **Pascal Titan X is 1.31x to 1.43x faster than the GTX 1080** and 1.47x to 1.60x faster than the Maxwell Titan X. This is without a doubt the **best card** you can get for deep learning right now.
- GTX 1080 > Maxwell Titan X: Across all models, the GTX 1080 is 1.10x to 1.15x faster than the Maxwell Titan X.
- ResNet > VGG: **ResNet-50 is 1.5x faster than VGG-16** and **more accurate than VGG-19** (7.02 vs 8.0); ResNet-101 is about the same speed as VGG-16 but much more accurate than VGG-19 (6.21 vs 8.0).
- **Always use cuDNN**: On the Pascal Titan X, cuDNN is 2.2x to 3.0x faster than nn; on the GTX 1080, cuDNN is 2.0x to 2.8x faster than nn; on the Maxwell Titan X, cuDNN is 2.2x to 3.0x faster than nn.
- GPUs are critical: The Pascal Titan X with cuDNN is **49x to 74x faster than dual Xeon E5-263**0 v3 CPUs.

All benchmarks were run in Torch. The GTX 1080 and Maxwell Titan X benchmarks were run on a machine with dual Intel Xeon E5-2630 v3 processors (8 cores each plus hyperthreading means 32 threads) and 64GB RAM running Ubuntu 14.04 with the CUDA 8.0 Release Candidate. The Pascal Titan X benchmarks were run on a machine with an Intel Core i5-6500 CPU and 16GB RAM running Ubuntu 16.04 with the CUDA 8.0 Release Candidate.

We benchmark all models with a **minibatch size of 16** and an image size of 224 x 224; this allows direct comparisons between models, and allows all but the ResNet-200 model to run on the GTX 1080, which has **only 8GB of memory**.

Multi-GPU Support ON FRAMEWORKS



http://deeplearning.net/software/theano/tutorial/using_multi_gpu.html

Training an Object Classifier in Torch-7 on multiple GPUs over ImageNet

In this concise example (1200 lines including a general-purpose and highly scalable data loader for images), we showcase:

- train AlexNet or Overfeat, VGG and Googlenet on ImageNet
- showcase multiple backends: CuDNN, CuNN
- use nn.DataParallelTable to speedup training over multiple GPUs
- multithreaded data-loading from disk (showcases sending tensors from one thread to another without serialization)

Requirements

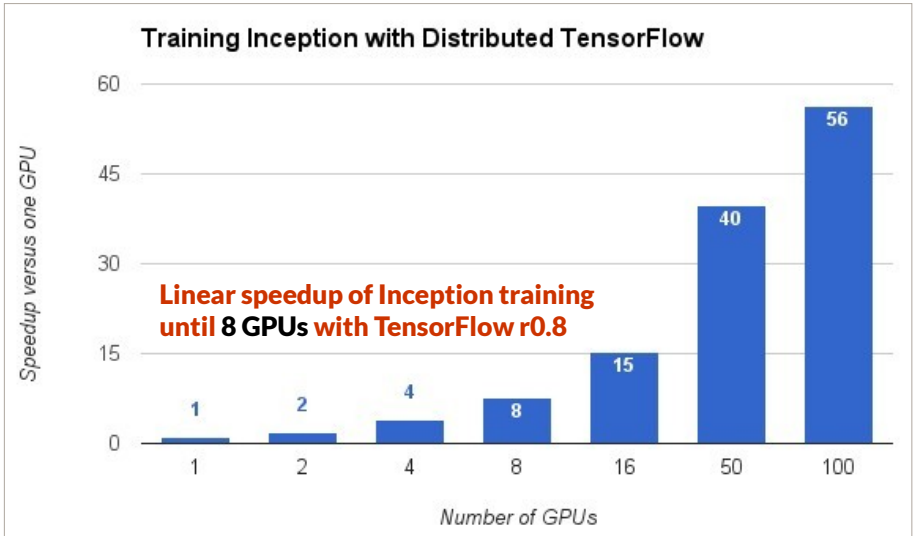
- Install torch on a machine with CUDA GPU
- If on Mac OSX, run `brew install coreutils findutils` to get GNU versions of `wc` , `find` , and `cut`
- Download Imagenet-12 dataset from http://image-net.org/download-images . It has 1000 classes and 1.2 million images.

https://github.com/soumith/imagenet-multiGPU.torch

Faster Training in neon with Multiple GPUs on the Nervana Cloud

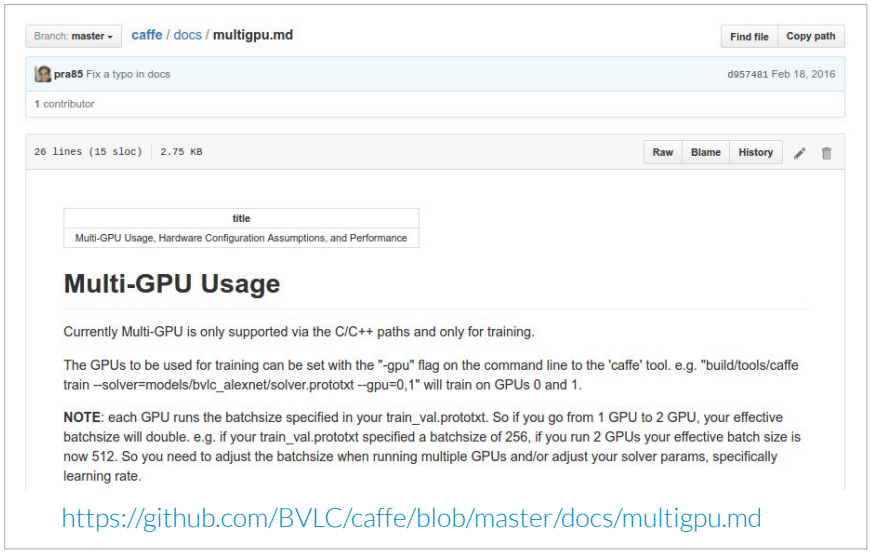
The Nervana Cloud provides unprecedented performance, ease of use, and the ability to apply deep learning to a large range of machine learning problems. With modern networks taking days, weeks or even months to train, performance is one of our fundamental goals. GPUs allow us to greatly improve performance by parallelizing convolution and matrix multiply operations over thousands of CUDA cores. Nervana tops the benchmarks for deep learning performance on single GPUs. In order to provide the best possible performance, we have extended this parallelization across multiple GPUs within a physical machine. Since the GPUs are linked together with high speed PCIe buses and direct peer-to-peer memory copies are possible through the driver, we have added a multi-GPU backend implemented with pycuda to our cloud version of neon.

nervanasys.com



TensorFlow can speed up Inception training by a factor of 56, using 100 GPUs.

research.googleblog.com



https://github.com/BVLC/caffe/blob/master/docs/multigpu.md

GPU WITH SSD HDD

the INQUIRER

Graphics

AMD launches £7,600 GPU with 1TB SSD storage attached

Fiji-based card features two PCIe 3.0 M.2 slots for adding NAND flash

AMD said that this makes the Radeon Pro SSG game-changing for 8K video, high-resolution rendering, VR content creation, oil and gas exploration, computational engineering, medical imaging and life sciences.

Applications that are willing to trade some performance (slower transfer from M.2 SSD compared to GDDR5) for massive available memory.

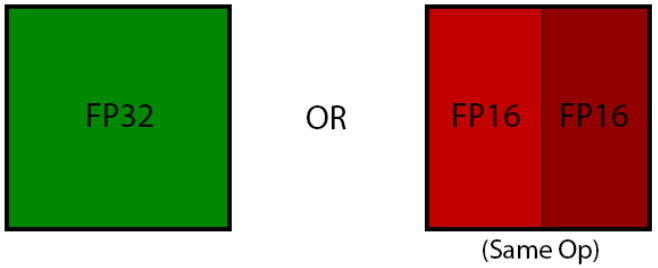
See the *"concept GPU"* from AMD. Note though the domination of NVIDIA in deep learning, but maybe NVIDIA will introduce a similar GPU.

Radeon Pro SSG, http://www.theinquirer.net/inquirer/news/2466103/amd-launches-gbp7-600-gpu-with-1tb-ssd-storage-attached

GPU: Training vs. Inference

- Possible to quantize (”compress”) the trained network from FP32 to FP16 with X1, but not with K1, and in theory double the inference speed

Tegra X1 CUDA Core FP Modes



NVIDIA Tegra GPU Specification Comparison		
	K1	X1
CUDA Cores	192	256
Texture Units	8	16
ROPs	4	16
GPU Clock	~950MHz	~1000MHz
Memory Clock	930MHz (LPDDR3)	1600MHz (LPDDR4)
Memory Bus Width	64-bit	64-bit
FP16 Peak	365 GFLOPS	1024 GFLOPS
FP32 Peak	365 GFLOPS	512 GFLOPS
Architecture	Kepler	Maxwell
Manufacturing Process	TSMC 28nm	TSMC 20nm SoC

http://www.anandtech.com/show/8811/nvidia-tegra-x1

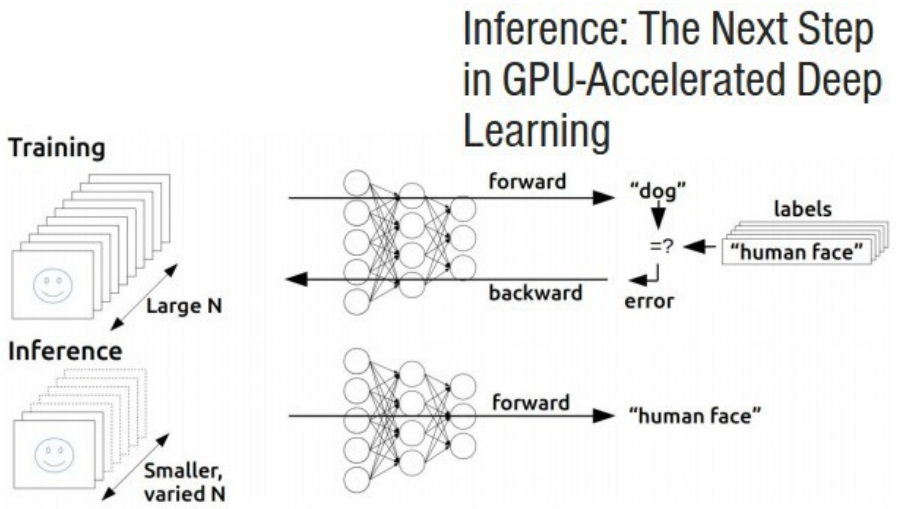


Figure 1: Deep learning training compared to inference. In training, many inputs, often in large batches, are used to train a deep neural network. In inference, the trained network is used to discover information within new inputs that are fed through the network in smaller batches.

Network: GoogLeNet	Batch Size	Titan X (FP32)	Tegra X1 (FP32)	Tegra X1 (FP16)
Inference Performance	1	138 img/sec	33 img/sec	33 img/sec
Power		119.0 W	5.0 W	4.0 W
Performance/Watt		1.2 img/sec/W	6.5 img/sec/W	8.3 img/sec/W
Inference Performance	128 (Titan X) 64 (Tegra X1)	863 img/sec	52 img/sec	75 img/sec
Power		225.0 W	5.9 W	5.8 W
Performance/Watt		3.8 img/sec/W	8.8 img/sec/W	12.8 img/sec/W

Table 3 GoogLeNet inference results on Tegra X1 and Titan X. Tegra X1's total memory capacity is not sufficient to run batch size 128 inference.

Note! Image size is 224x224x3 (RGB) for GoogleNet
→ (224 x 224 x 3 x 16 bits x (1/8 byte/bits) =294kB image itself

https://www.nvidia.com/content/tegra
https://devblogs.nvidia.com/parallelforall/inference-next-step

Optimizing deployed networks

Final trained networks may contain:

- Redundant weights
- 'Too high precision' for inference, GPU with **int8 inference** could be even useful

SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer

(Submitted on 24 Feb 2016 (v1), last revised 6 Apr 2016 (this version, v3))

Introduction and Motivation

More efficient distributed training. Communication among servers is the limiting factor to the scalability of distributed CNN training. For distributed data-parallel training, communication overhead is directly proportional to the number of parameters in the model [15]. In short, small models train faster due to requiring less communication.

Less overhead when exporting new models to clients. For autonomous driving, companies such as Tesla periodically copy new models from their servers to customers' cars. With AlexNet. this would require 240MB of communication from the server to the car. Smaller models require less communication, making frequent updates more feasible.

Feasible FPGA and embedded deployment. FPGAs often have less than 10MB[1] of on-chip memory and no off-chip memory or storage. For inference, a sufficiently small model could be stored directly on the FPGA instead of being bottle necked by memory bandwidth [25], while video frames stream through the FPGA in real time.

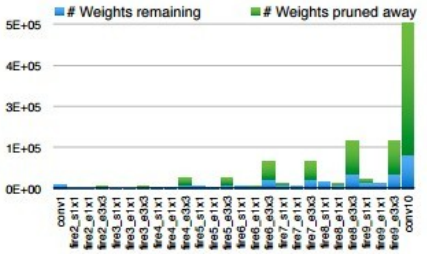


Figure 5. Parameters pruned away in each layer. 3x3 kernels has more redundancy and could be pruned more than 1x1 kernel. Overall SqueezeNet can be removed 3x parameters without hurting accuracy.

"We have presented SqueezeNet, a CNN architecture that has 50× fewer parameters than AlexNet and maintains AlexNet-level accuracy on ImageNet. We also compressed SqueezeNet to less than 0.5MB, or 510× smaller than AlexNet without compression."

Fixed Point Quantization of Deep Convolutional Networks

Darryl D. Lin DARRYL.DLIN@GMAIL.COM
Qualcomm Research, San Diego, CA 92121, USA
Sachin S. Talathi TALATHI@GMAIL.COM
Qualcomm Research, San Diego, CA 92121, USA
V. Sreekanth Annapureddy SREEKANTHAV@GMAIL.COM
NetraDyne Inc., San Diego, CA 92121, USA

http://arxiv.org/abs/1511.06393

Table 7. CIFAR-10 classification error rate with different bit-width combinations

Activation Bit-width	Weight Bit-width 4	8	16	Float
4	8.30	7.50	7.40	7.44
8	7.58	6.95	6.95	6.78
16	7.58	6.82	6.92	6.83
Float	7.62	6.94	6.96	6.98

Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications

Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, Dongjun Shin

(Submitted on 20 Nov 2015 (v1), last revised 24 Feb 2016 (this version, v2))

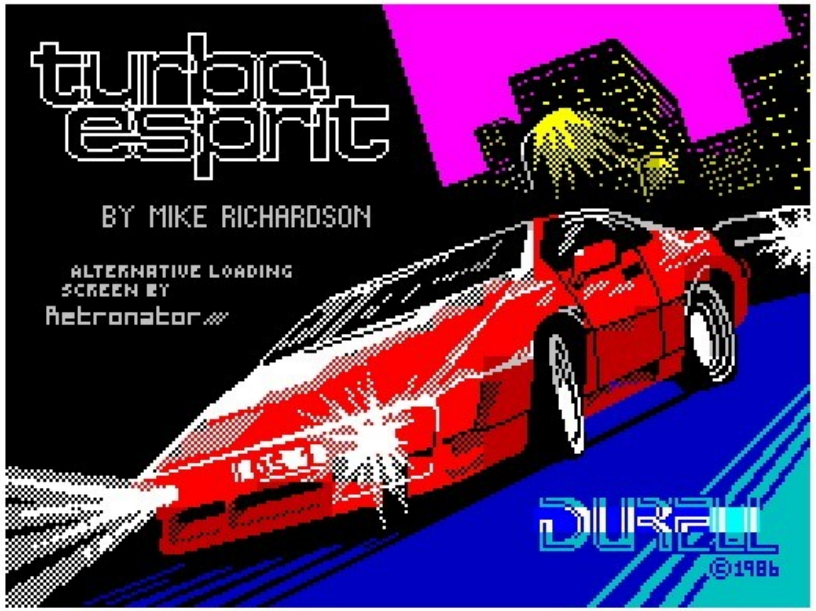
Table 1: Original versus compressed CNNs. Memory, runtime and energy are significantly reduced with only minor accuracy drop. We report the time and energy consumption for processing single image in S6 and Titan X. (* compression, S6: Samsung Galaxy S6).

Model	Top-5	Weights	FLOPs	S6		Titan X
AlexNet	80.03	61M	725M	117ms	245mJ	0.54ms
*AlexNet**	78.33	11M	272M	43ms	72mJ	0.30ms
(imp.)	(-1.70)	(×5.46)	(×2.67)	(×2.72)	(×3.41)	(×1.81)
VGG-S	84.60	103M	2640M	357ms	825mJ	1.86ms
*VGG-S**	84.05	14M	549M	97ms	193mJ	0.92ms
(imp.)	(-0.55)	(×7.40)	(×4.80)	(×3.68)	(×4.26)	(×2.01)
GoogLeNet	88.90	6.9M	1566M	273ms	473mJ	1.83ms
*GoogLeNet**	88.66	4.7M	760M	192ms	296mJ	1.48ms
(imp.)	(-0.24)	(×1.28)	(×2.06)	(×1.42)	(×1.60)	(×1.23)
VGG-16	89.90	138M	15484M	1926ms	4757mJ	10.67ms
*VGG-16**	89.40	127M	3139M	576ms	1346mJ	4.58ms
(imp.)	(-0.50)	(×1.09)	(×4.93)	(×3.34)	(×3.53)	(×2.33)

http://arxiv.org/abs/1511.06530

FIXED-POINT networks

Why are Eight Bits Enough for Deep Neural Networks?



https://petewarden.com/2015/05/23/

DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients

Shuchang Zhou , Zekun Ni , Xinyu Zhou , He Wen , Yuxin Wu , Yuheng Zou
Megvii Inc.
{shuchang.zhou,mike.zekun}@gmail.com
{zxy,wenhe,wyx,zouyuheng}@megvii.com

http://arxiv.org/abs/1606.06160

Training deep neural networks with low precision multiplications

Matthieu Courbariaux, Yoshua Bengio, Jean-Pierre David

(Submitted on 22 Dec 2014 (v1), last revised 23 Sep 2015 (this version, v5))

Improving the speed of neural networks on CPUs

Vincent Vanhoucke
Google, Inc.
Mountain View, CA 94043
vanhoucke@google.com

Andrew Senior
Google, Inc.
New York, NY 10011
andrewsenior@google.com

Mark Z. Mao
Google, Inc.
Mountain View, CA 94043
markmao@google.com

Hardware-oriented Approximation of Convolutional Neural Networks

Philipp Gysel, Mohammad Motamedi & Soheil Ghiasi
Department of Electrical and Computer Engineering
University of California, Davis
Davis, CA 95616, USA
{pmgysel,mmotamedi,ghiasi}@ucdavis.edu

Table 1: Quantization results for different parts of three networks. Only one number category is cast to fixed point, and the remaining numbers are in floating point format.

Fixed point bit-width	16-bit	8-bit	4-bit	2-bit
LeNet, 32-bit floating point accuracy: 99.1%				
Layer output	99.1%	99.1%	98.9%	85.9%
CONV parameters	99.1%	99.1%	99.1%	98.9%
FC parameters	99.1%	99.1%	98.9%	98.7%
Full CIFAR-10, 32-bit floating point accuracy: 81.7%				
Layer output	81.6%	81.6%	79.6%	48.0%
CONV parameters	81.7%	81.4%	75.9%	19.1%
FC parameters	81.7%	80.8%	79.9%	77.5%
CaffeNet top-1, 32-bit floating point accuracy: 56.9%				
Layer output	56.8%	56.7%	06.0%	00.1%
CONV parameters	56.9%	56.7%	00.1%	00.1%
FC parameters	56.9%	56.3%	00.1%	00.1%

Table 2: Fine-tuned networks with dynamic fixed point parameters and outputs for convolutional and fully connected layers. The numbers in brackets indicate accuracy without fine-tuning.

	Layer outputs	CONV parameters	FC parameters	32-bit floating point baseline	Fixed point accuracy
LeNet (Exp 1)	4-bit	4-bit	4-bit	99.1%	99.0% (98.7%)
LeNet (Exp 2)	4-bit	2-bit	2-bit	99.1%	98.8% (98.0%)
Full CIFAR-10	8-bit	8-bit	8-bit	81.7%	81.4% (80.6%)
SqueezeNet top-1	8-bit	8-bit	8-bit	57.7%	57.1% (55.2%)
CaffeNet top-1	8-bit	8-bit	8-bit	56.9%	56.0% (55.8%)
GoogLeNet top-1	8-bit	8-bit	8-bit	68.9%	66.6% (66.1%)

GPU on the go (Embedded GPU)

- If you need to deploy your network in a more portable form
  - e.g. autonomous car, drones, portable medical imaging, etc.
- Train in desktop/server, and do the inference on the embedded NVIDIA chip, Jetson (Tegra) TK1/TX1 for example

JETSON TK1 **0.3 TFLOPS**

Tegra K1 SOC

- NVIDIA Kepler GPU with 192 CUDA Cores
- NVIDIA 4-Plus-1™ Quad-Core ARM® Cortex™-A15 CPU

Our Price: £132.22

Nvidia offers a version of Linux based on Ubuntu 14.04 for the board, and most ARM-based games are for mobile devices and written for the Android OS.

http://www.nvidia.com/object/jetson-tk1-dev

JETSON TX1 **1 TFLOPS**

NVIDIA

$599

BUY NOW

JETSON TX1 MODULE

- NVIDIA Maxwell™ GPU with 256 NVIDIA® CUDA® Cores
- Quad-core ARM® Cortex®-A57 MPCore Processor

It comes pre-flashed with a Linux environment, includes support for many common APIs, and is supported by NVIDIA's complete development tool chain. The board also exposes a variety of standard hardware interfaces, enabling a highly flexible and extensible platform. This makes it ideal for all your applications requiring high computational performance in a low-power envelope.

http://www.nvidia.com/object/jetson-tx1-dev

Jetson TX1 FRAMEWORKS

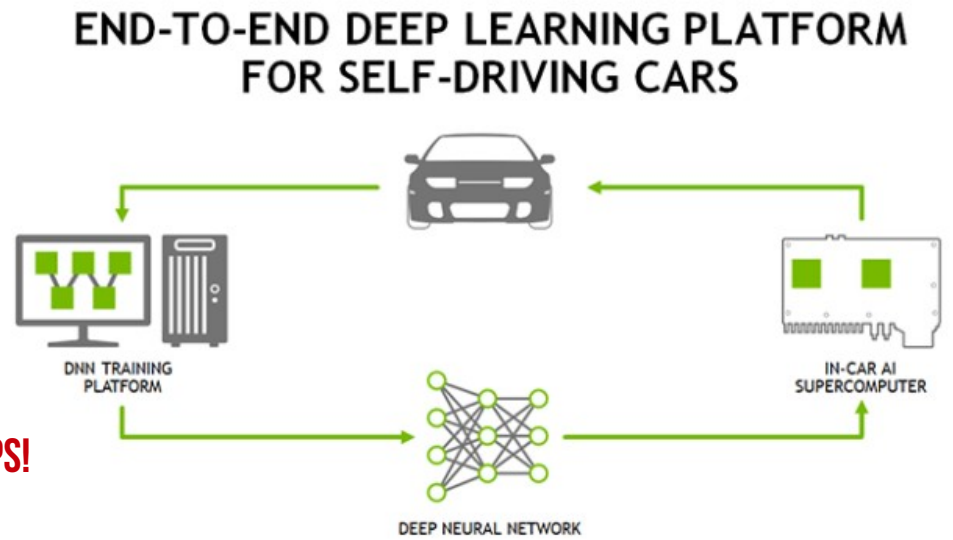
- If you plan to train on workstation, and do the inference on the Jetson TX1, note that at the time of the writing (Aug, 2016) some frameworks seem to be easier to install than others.
  - **Theano**: Easy Installation of an Optimized Theano on Current Ubuntu
  - **Caffe**: CaffePresso: An Optimized Library for Deep Learning on Embedded Accelerator-based platforms, JetsonHacks - Caffe Deep Learning Framework - NVIDIA Jetson TX1
  - **TensorFlow**: Could be slightly difficult to make work, https://github.com/tensorflow/tensorflow/issues/851
  - **Torch**: Older TK1 works, as well as the TX1. https://github.com/dusty-nv/rovernet/blob/master/CMakePreBuild.sh
  - **Nervana Neon**: Should work, https://github.com/NervanaSystems/neon/issues/175

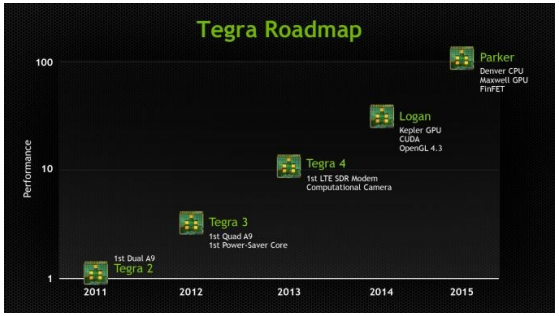
Next-gen NVIDIA Tegra

NVIDIA DRIVE PX Specification Comparison		
	DRIVE PX	DRIVE PX 2
SoCs	2x Tegra X1	2x Tegra "Parker"
Discrete GPUs	N/A	2x Unknown Pascal
CPU Cores	8x ARM Cortex-A57 + 8x ARM Cortex-53	4x NVIDIA Denver + 8x ARM Cortex-A57
GPU Cores	2x Tegra X1 (Maxwell)	2x Tegra "Parker" (Pascal) + 2x Unknown Pascal
FP32 TFLOPS	> 1 TFLOPS	8 TFLOPS
FP16 TFLOPS	> 2 TFLOPS	16 TFLOPS?
TDP	N/A	250W

MORE FLOPS!

http://www.anandtech.com/

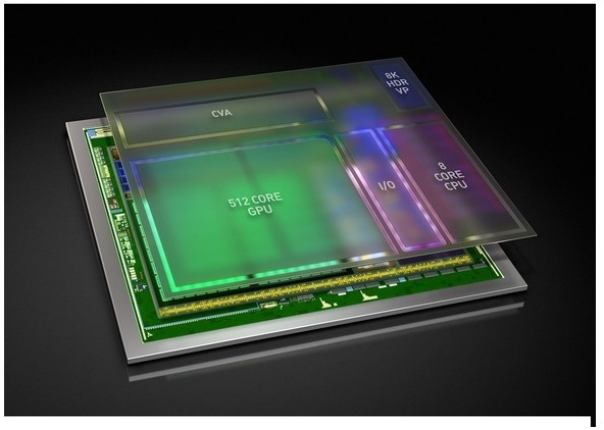




http://www.anandtech.com/

Nvidia teases Volta GPU in next-gen Xavier self-driving car computer

Nvidia's upcoming 512-core Volta GPU will help next-generation Xavier supercomputer chip make self-driving cars safer



http://www.pcworld.com/article/3125456/

Xavier delivers 20 TOPS (trillion operations per second) of computing power while consuming only 20 watts of power. It has seven billion transistors and is made using the 16-nm chip process. Those specs hint big gains in performance and power efficiency for the Volta GPU.

"A single **Xavier AI processor will be able to replace today's Drive PX 2** configured with dual mobile SoCs and dual discrete GPUs—at a fraction of the power consumption," Nvidia said.

The Drive PX 2, which was introduced at CES in January, has 12 CPU cores and two Pascal GPUs. It was the equivalent of having "150 MacBook Pros in your trunk," said Nvidia CEO Jen-Hsun Huang.

“Exotic” Inference

- Compared to “NVIDIA-dominated” training, there are some more options for inference for client-side inference.

Deep Learning On A Stick: Movidius'

http://www.tomshardware.com/news/movidius-fathom-neural-compute-stick,31694.html

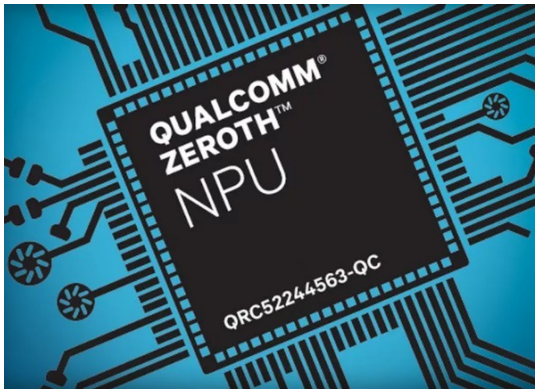
The Fathom's performance ranges from **80 to 150 GFLOPS**, depending on the neural network's complexity and precision (**8-bit** and 16-bit precision is supported). That performance requires less than **1.2W** of power, which is 12x lower power than, for example, what an Nvidia Jetson TX1

Target price? Under $100 (about £70 or AU$130).

Qualcomm's deep learning SDK will mean more AI on your smartphone

Deep learning used to only be in the cloud — now it's coming to devices

By James Vincent on May 2, 2016 08:25 am @jvincent



Qualcomm's Zeroth SDK (called the Qualcomm Snapdragon Neural Processing Engine) gives manufacturers and companies an easy toolset to run limited deep learning programs locally with Snapdragon 820 processors.
theverge.com

tom's HARDWARE
THE AUTHORITY ON TECH

What are y

TAGS: Builds Cases Cooling CPUs Gaming

CHIPSETS > NEWS

The Rise Of Client-Side Deep Learning

by Lucian Armasu May 12, 2016 at 6:48 PM

Forbes

SEP 8, 2016 @ 01:40 PM 994 VIEWS The Little Black Book of Billionaire Secrets

Intel Boosts Its Perceptual Computing Strategy With The Movidius Acquisition

http://www.forbes.com/sites/greatspeculations/2016/09/08

DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices

Nicholas D. Lane‡, Sourav Bhattacharya‡, Petko Georgiev† Claudio Forlivesi‡, Lei Jiao‡, Lorena Qendro∗, and Fahim Kawsar‡ ‡Bell Labs, †University of Cambridge, ∗University of Bologna

http://dx.doi.org/10.1109/IPSN.2016.7460664

Nic Lane, UCL: “Deep Learning for Embedded Devices: Next Step in Privacy-Preserving High-Precision Mobile Health and Wellbeing Tools” at Deep Learning Summit 2016, London

Neuromorphic Chips Future?

IBM's Brain-Inspired Chip Tested for Deep Learning

IEEE SPECTRUM

By Jeremy Hsu
Posted 27 Sep 2016 | 20:00 GMT

http://spectrum.ieee.org

Illustration: Shutterstock

IBM published a paper on its work in the 9 Sept 2016 issue of the journal Proceedings of the National Academy of Sciences. The research was funded with just under $1 million from the U.S. Defense Advanced Research Projects Agency (DARPA). Such funding formed part of DARPA's Cortical Processor program aimed at brain-inspired AI that can recognize complex patterns and adapt to changing environments.

Yann LeCun, director of AI research at Facebook and a pioneer in deep learning, previously critiqued IBM's TrueNorth chip because it primarily supports spiking neural networks. (See IEEE Spectrum's previous interview with LeCun on deep learning.)

The IBM TrueNorth design may better support the goals of neuromorphic computing that focus on closely mimicking and understanding biological brains, says Zachary Chase Lipton, a deep-learning researcher in the Artificial Intelligence Group. at the University of California, San Diego.

Such biologically-inspired chips would probably become popular only if they show that they can outperform other hardware approaches for deep learning, Lipton says. But he suggested that IBM could leverage its hardware expertise to join Google and Intel in creating new specialized chips designed specifically for deep learning.

Proceedings of the National Academy of Sciences of the United States of America PNAS

> Early Edition > Steven K. Esser, doi: 10.1073/pnas.1604850113

CrossMark click for updates

Convolutional networks for fast, energy-efficient neuromorphic computing

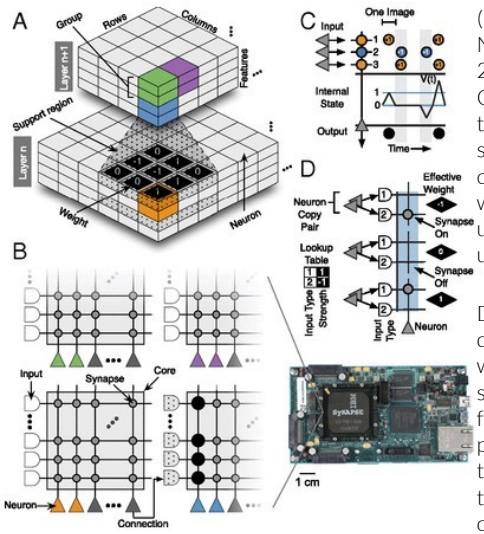
Steven K. Esser[a,1], Paul A. Merolla[a], John V. Arthur[a], Andrew S. Cassidy[a], Rathinakumar Appuswamy[a], Alexander Andreopoulos[a], David J. Berg[a], Jeffrey L. McKinstry[a], Timothy Melano[a], Davis R. Barch[a], Carmelo di Nolfo[a], Pallab Datta[a], Arnon Amir[a], Brian Taba[a], Myron D. Flickner[a], and Dharmendra S. Modha[a]

Author Affiliations [a]Brain-Inspired Computing, IBM Research–Almaden, San Jose, CA 95120

Edited by Karlheinz Meier, University of Heidelberg, Heidelberg, Germany, and accepted by Editorial Board Member J. A. Movshon August 9, 2016 (received for review March 24, 2016)

http://dx.doi.org/10.1073/pnas.1604850113

Inference on TrueNorth



(B) A TrueNorth chip (shown far right socketed in IBM's NS1e board) comprises 4,096 cores, each with 256 inputs, 256 neurons, and a 256 ×× 256 synaptic array. Convolutional network neurons for one group at one topographic location are implemented using neurons on the same TrueNorth core (TrueNorth neuron colors correspond to convolutional network neuron colors in A), with their corresponding filter support region implemented using the core's inputs, and filter weights implemented using the core's synaptic array.

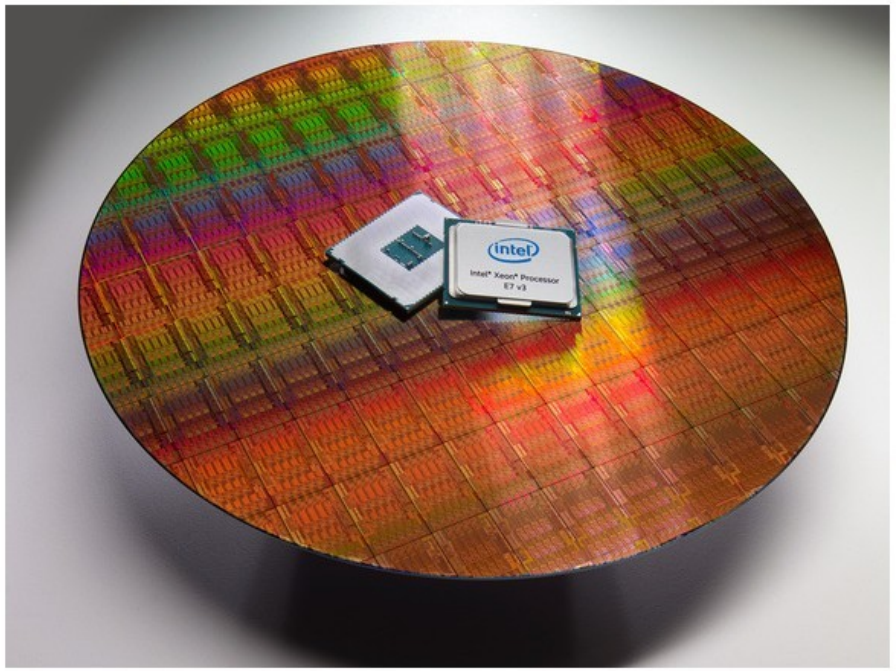
D) Convolutional network filter weights (numbers in black diamonds) implemented using TrueNorth, which supports weights with individually configured on/off state and strength assigned by lookup table. In our scheme, each feature is represented with pairs of neuron copies. Each pair connects to two inputs on the same target core, with the inputs assigned types 1 and 2, which via the look up table assign strengths of +1+1 or −1−1 to synapses on the corresponding input lines. By turning on the appropriate synapses, each synapse pair can be used to represent −1−1, 0, or +1+1.

CPU i7 vs Xeon

During actual deep learning training and inference, the CPU choice is not very crucial but in general one needs to do computations outside GPU as well for example when doing data wrangling to get data “trainable”.

40 PCIe lanes should be found from CPU to use your GPUs as efficient as possible. Xeon supports a lot more RAM.

Intel's 18-core Xeon chips tuned for machine learning, analytics



PCWORLD.COM

Xeon vs i7/i5 – What’s the difference?

VELOCITYMICRO.COM

CPU Features

Looking at just the specifications, it would appear that there is not much that differentiates a Core i7 and Xeon CPU. This is because the main differences are not spec-based, but rather the features found in each product line:

Product Line	Overclocking Support	Max CPUs	Max Memory	ECC RAM Support	VPro	VT-x/VT-d	TXT
High End Core i7	Yes	1	64GB	No	No	Yes	No
Single Socket Xeon E5 v3	No	1	768 GB	Yes	Yes	Yes	Yes
Dual Socket Xeon E5 v3	No	2	768 GB	Yes	Yes	Yes	Yes

PUGETSYSTEMS.COM

CPU i7 SINGLE CPU 40 LANES → DUAL-CPU 80 LANES

The New CPUs

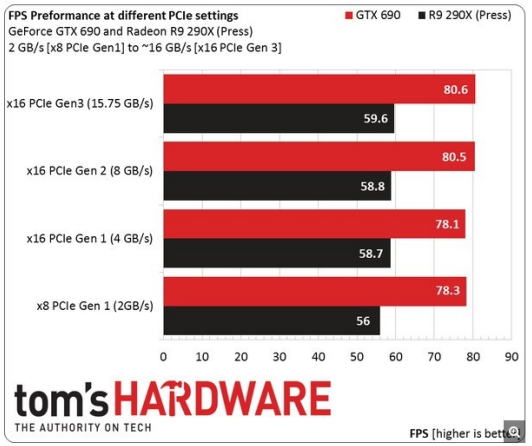
Getting straight to the heart of the matter, Intel is keeping the enthusiast extreme range simple by only releasing three models, similar to the initial Sandy Bridge-E and Ivy Bridge-E launches.

Brand Name & Processor Number[1]	Base Clock Speed (GHz)	Turbo Frequency[2] (GHz)	Cores/ Threads	Cache	PCI Express* 3.0 Lanes	Memory Support	TDP	Socket (LGA)	Pricing (1K USD)
NEW Intel® Core™ i7 5960X	3.0	Up to 3.5	8/16	20MB	40	4 channels DDR4-2133	140W	2011-v3	$999
NEW Intel® Core™ i7 5930K	3.5	Up to 3.7	6/12	15MB	40	4 channels DDR4-2133	140W	2011-v3	$583
NEW Intel® Core™ i7 5820K	3.3	Up to 3.6	6/12	15MB	28	4 channels DDR4-2133	140W	2011-v3	$389
Intel® Core™ i7 4790K	4.0	Up to 4.4	4/8	8MB	16	2 channels DDR3-1600	88W	1150	$339
Intel® Core™ i5 4690K	3.5	Up to 3.9	4/4	6MB	16	2 channels DDR3-1600	88W	1150	$242

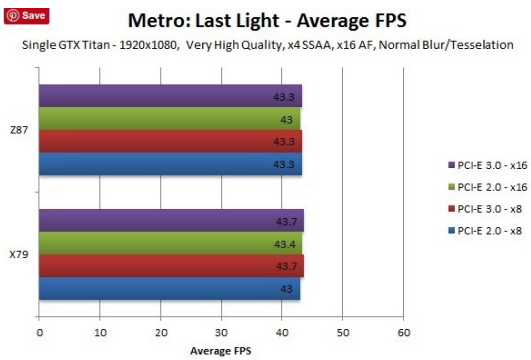
The top of the line will be the 8-core i7-5960X with HyperThreading, using a 3.0 GHz base frequency and 40 PCIe 3.0 lanes for $999 for 1000 units. This pricing is in line with previous extreme edition processor launches, but the base frequency is quite low. This is due to the TDP limitation: sticking two extra cores produces extra energy lost as heat, and in order to get the TDP down the base clock has to be reduced over the six-core models. This is a common trend we see in the Xeon range, and as a result it might affect the feel of day-to-day performance.

http://www.anandtech.com/show/8426/the-intel-haswell-e-cpu-review-core-i7-5960x-i7-5930k-i7-5820k-tested

IN THEORY, FASTER TO RUN EVERY GPU AT X16, BUT IN PRACTICE, IS THE 1 X16/3 X8 THAT BAD?



tomshardware.com, older generation gaming GPUs



pugetsystems.com GTX Titan 6GB in gaming

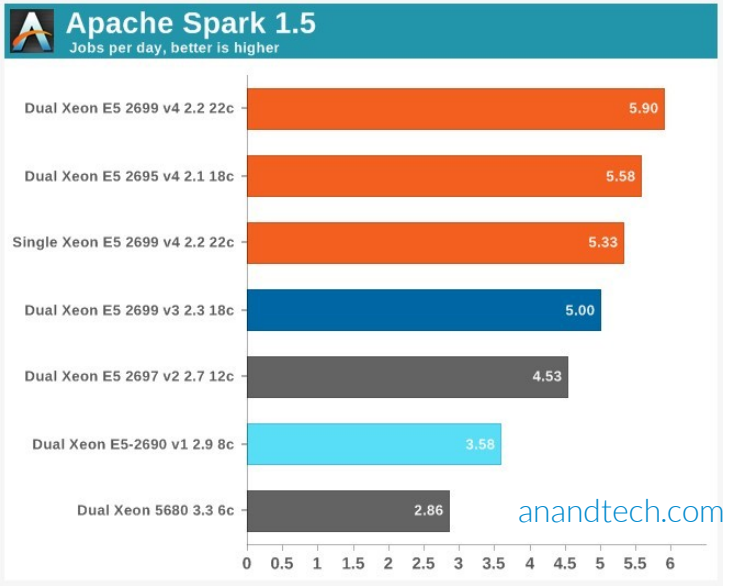
DEEP LEARNING BENCHMARKS?

CPU Xeon E5-26xx "2" FOR DOUBLE-CPU SYSTEMS

Intel Xeon E5-2600 v4 Series

Feature	Intel® Xeon® processor E5-2600 v3 product family (HSW-EP)	Intel® Xeon® processor E5-2600 v4 product family (BDW-EP)
Socket	Socket R3	
Process Technology	22 nm	14 nm
Architecture	Haswell microarchitecture	Broadwell microarchitecture
Max Core/Thread Count	Up to 18 cores / 36 threads	Up to 22 cores / 44 threads
Memory Support	4xDDR4 channels	
Memory Speed	Up to 2133 MT/s	Up to 2400 MT/s
Max: Memory Channels / DIMM Slots / Capacity	8 / 24 / 1536 GB	
QPI Ports	2x QPI 1.1 channels 6.4, 8.0, 9.6 GT/s	
PCIe* Lanes / Controllers	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	
TDP (W)	Up to 145W Server; 160W Workstation only	
Chipset	Intel® C610 series chipset (Wellsburg PCH)	
Connectivity	Up to 40GbE - Intel® Ethernet Controller XL710 (Fortville)	

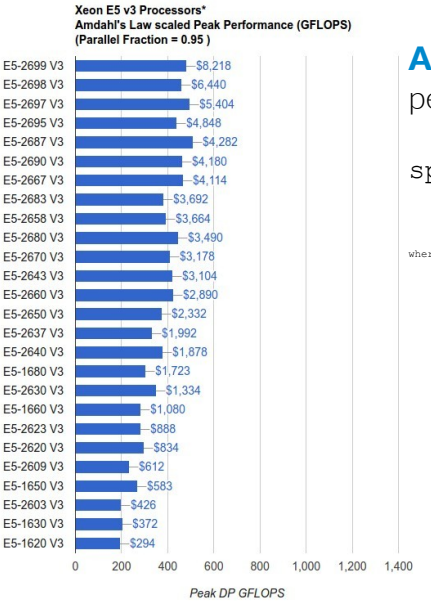
tomshardware.co.uk



anandtech.com

Spark threw us back into nineties, to the time that several workloads still took ages on high-end computers. It takes no less than six and half hours on a 16-core Xeon E5-2690 running at 2.9 GHz to crunch through 300 GB of web data and extract anything meaningful out of it. So we have to express our times in "jobs per day"

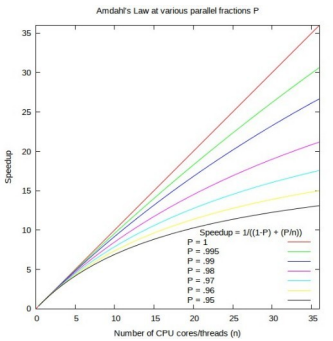
newegg.com



Amdahl's law limits the multithread performance in practice

```
speedup = 1/( (1-P) + P/n )
```

where P is the parallel fraction and n is the number of processes (cores)



CPU Xeon E5-26xx V3 VS V4

Product Name	Intel® Xeon® Processor E5-2620 v3 (15M Cache, 2.40 GHz)	Intel® Xeon® Processor E5-2620 v4 (20M Cache, 2.10 GHz)
Code Name	Haswell	Broadwell
Essentials		
Status	Launched	Launched
Launch Date	Q3'14	Q1'16
Processor Number	E5-2620V3	E5-2620V4
Cache	15 MB SmartCache	20 MB SmartCache
Performance		
# of Cores	6	8
# of Threads	12	16
Processor Base Frequency	2.4 GHz	2.1 GHz
Max Turbo Frequency	3.2 GHz	3 GHz
TDP	85 W	85 W
Memory Specifications		
Max Memory Size (dependent on memory type)	768 GB	1536 GB
Memory Types	DDR4 1600/1866	DDR4 1600/1866/2133
Max # of Memory Channels	4	4
Max Memory Bandwidth	59 GB/s	68.3 GB/s
Physical Address Extensions	46-bit	46-bit
ECC Memory Supported ‡	Yes	Yes
Expansion Options		
PCI Express Revision	3.0	3.0
PCI Express Configurations ‡	x4, x8, x16	x4, x8, x16
Max # of PCI Express Lanes	40	40
Package Specifications		
Max CPU Configuration	2	2
T_{CASE}	72.6°C	74°C
Package Size	52.5mm x 45mm	45mm x 52.5mm
Sockets Supported	FCLGA2011-3	FCLGA2011-3
Low Halogen Options Available	See MDDS	See MDDS

Pros and Cons summary

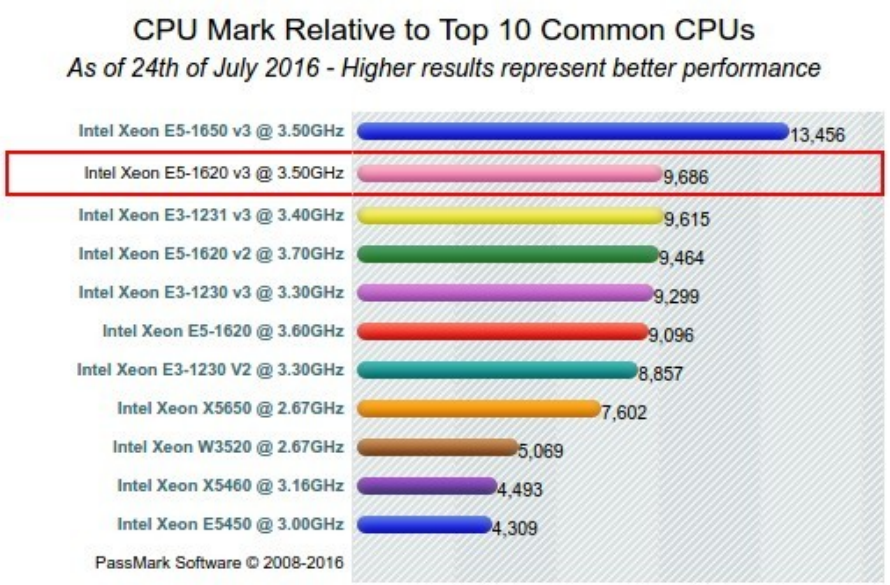
Core i7-5930K	Xeon E5-2620 v4
General recommendations: Clocked higher, Can be easily overclocked	**General recommendations:** Supports dual-processing, More cores for better multi-threading performance, Can execute more threads at once, Needs less power, Somewhat lower official price
Drawbacks: Not capable of dual-processing, Doesn't have as many CPU cores, Executes fewer threads, Requires much more power, Somewhat higher official price	**Drawbacks:** Clocked lower

Core i7 5930K vs Intel Xeon E5-2620 v4

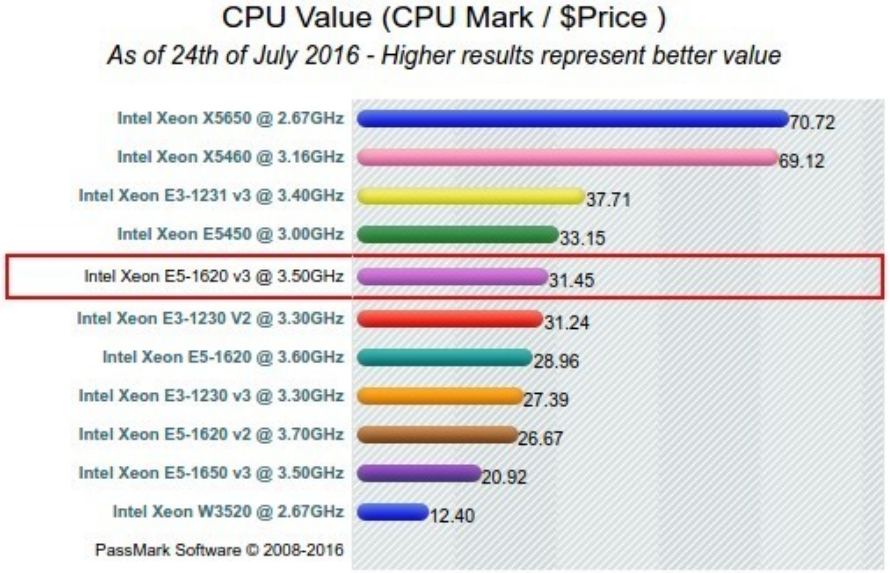
Dual CPU system with Haswell-EP (22nm) **E5-26xx** gives good multicore performance for relatively low price, and allows **a lot more RAM** than the i7 systems.

It is especially useful for **“parfor”** pre-processing of datasets with Matlab (single thread for file for example). If you don't have use for this, you might find the i7 line more suitable with better single core performance

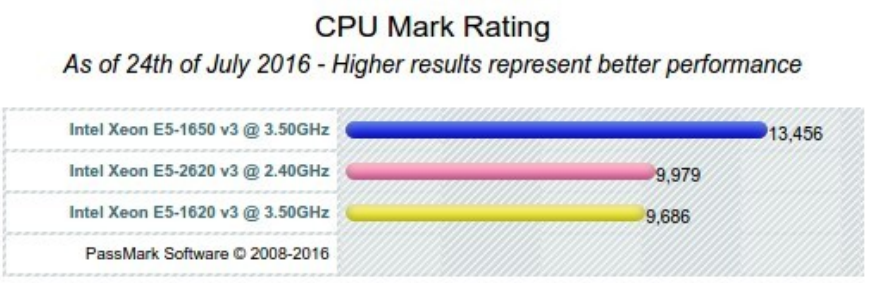
CPU Xeon E5-16xx "1" DO NOT WORK IN DUAL-CPU CONFIG

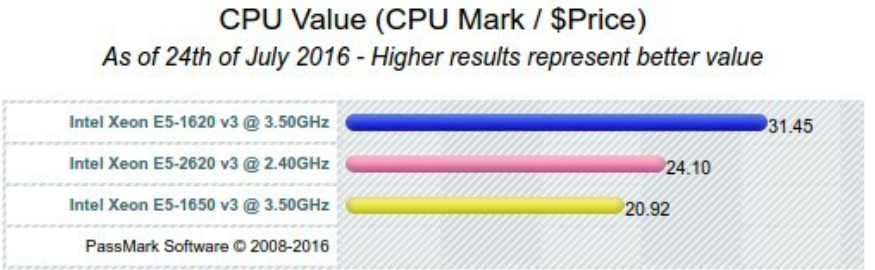


	Intel Xeon E5-1620 v3 @ 3.50GHz	Intel Xeon E5-1650 v3 @ 3.50GHz	Intel Xeon E5-2620 v3 @ 2.40GHz
Price	$307.99 BUY NOW!	$643.15 BUY NOW!	$413.99 BUY NOW!
Socket Type	LGA2011-v3	LGA2011-v3	LGA2011-v3
CPU Class	Server	Server	Server
Clockspeed	3.5 GHz	3.5 GHz	2.4 GHz
Turbo Speed	Up to 3.6 GHz	Up to 3.8 GHz	Up to 3.2 GHz
# of Physical Cores	4 (2 logical cores per physical)	6 (2 logical cores per physical)	6 (2 logical cores per physical)
Max TDP	140W	140W	85W
First Seen on Chart	Q2 2014	Q4 2014	Q4 2014
# of Samples	181	224	60
Single Thread Rating	2005	2114	1692
CPU Mark	9686	13456	9979
	£274.56	£498.49	£391.49



http://www.cpubenchmark.net/cpu.php?cpu=Intel+Xeon+E5-1620+v3+%40+3.50GHz





http://www.cpubenchmark.net

CPU Xeon E5-1650 V3 vs V4

At least the switch from E5-26xxV3 to E5-26xxV4 (14-core also priced at £2,000) seems to be bringing some relative gains to machine learning and number crunching applications (anandtech.com), but could not find similar info on E5-16xx. Minor improvements over v3 for sure, the significance in practice?

If the price difference is not very significant, you could go for the V4 instead of the older V3.

	v3	v4
Performance		
# of Cores	6	6
# of Threads	12	12
Processor Base Frequency	3.5 GHz	3.6 GHz
Max Turbo Frequency	3.8 GHz	4 GHz
TDP	140 W	140 W

	v3	v4
Memory Specifications		
Max Memory Size (dependent on memory type)	768 GB	1536 GB
Memory Types	DDR4 1333/1600/1866/2133	DDR4 1600/1866/2133/2400
Max # of Memory Channels	4	4
Max Memory Bandwidth	68 GB/s	76.8 GB/s
Physical Address Extensions	46-bit	46-bit
ECC Memory Supported ‡	Yes	Yes

http://ark.intel.com/compare/82765,92994

	Intel Xeon E5-1650 v4 @ 3.60GHz	Intel Xeon E5-1650 v3 @ 3.50GHz
Price	Search Online	$644.38 BUY NOW!
Socket Type	FCLGA2011-3	LGA2011-v3
CPU Class	Server	Server
Clockspeed	3.6 GHz	3.5 GHz
Turbo Speed	Up to 4.0 GHz	Up to 3.8 GHz
# of Physical Cores	6 (2 logical cores per physical)	6 (2 logical cores per physical)
Max TDP	140W	140W
First Seen on Chart	Q2 2016	Q4 2014
# of Samples	2	226
Single Thread Rating	2025	2115
CPU Mark	13509	13460

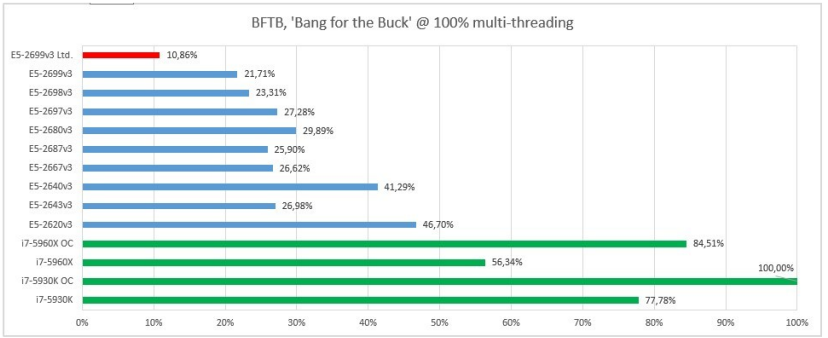
cpubenchmark.net

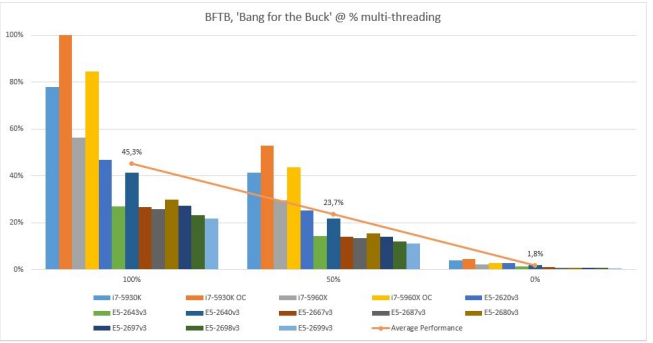
	Intel Xeon E5-1650 v3	Intel Xeon E5-1650 v4
Market segment	Server	
Manufacturer	Intel	
Family	Xeon	
Basic details		
Model number	E5-1650 v3	E5-1650 v4
CPU part number	CM8064401548111	CM8066002044306
Box part number	BX80644E51650V3	BX80660E51650V4
Introduction date	September 8, 2014	June 7, 2016
Current official price	$583 (as of Jul 2016)	$617 (as of Jul 2016)
CPU features		
Core name	Haswell-EP	Broadwell-EP
Microarchitecture	Haswell	Broadwell
Technology (micron)	0.022	0.014
Data width (bits)	64	
Socket	Socket 2011-3	
Frequency (MHz)	3500	3600
Turbo Frequency (MHz)	3800	4000
Clock Multiplier	35	36
L1 cache	192 KB (code) / 192 KB (data)	
L2 cache (KB)	1536	
L3 cache (KB)	15360	
Max temperature (°C)	66.7	69
TDP (Watt)	140	
Core voltage (V)	0.65 - 1.3	
Cores	6	
Threads	12	
Multiprocessing	1	
Instruction set extensions		
AES / Advanced Encryption Standard	+	
AMD64 / EM64T 64-bit technology	+	
AVX / Advanced Vector Extensions	+	
AVX2 / Advanced Vector Extensions 2	+	
F16C / 16-bit Floating-Point conversion	+	
FMA3 / 3-operand Fused Multiply-Add	+	
MMX	+	
SSE	+	
SSE2	+	
SSE3	+	
SSE4.1	+	
SSE4.2	+	
SSSE3 / Supplemental SSE3	+	
TSX / Transactional Synchronization Extensions	-	+
Supported technologies		
Hyper-Threading	+	
PowerNow! / Enhanced SpeedStep	+	
Trusted Execution	+	
Turbo Core / Turbo Boost	+	
Virtualization	+	
Virus Protection / Execute Disable bit	+	
Integrated Graphics		
GPU Type	None	
Integrated Memory Controller(s)		
The number of controllers	1	
Memory channels	4	
Supported memory	DDR4-1333, DDR4-1600, DDR4-1866, DDR4-2133	DDR4-1600, DDR4-1866, DDR4-2133, DDR4-2400
Maximum memory bandwidth (GB/s)	68	76.8
ECC supported	Yes	

cpu-world.com

Single-thread vs. Multi-thread

- Multi-thread performance is not the same as single-thread performance multiplied with number of cores.
  - Depends how parallel are your algorithm, and what is the associated overhead (see Amdahl's law in previous E5-26xx slides).
  - Hard to interpret benchmarking results as most of them are typically work gaming and “normal tasks” rather than for machine learning.

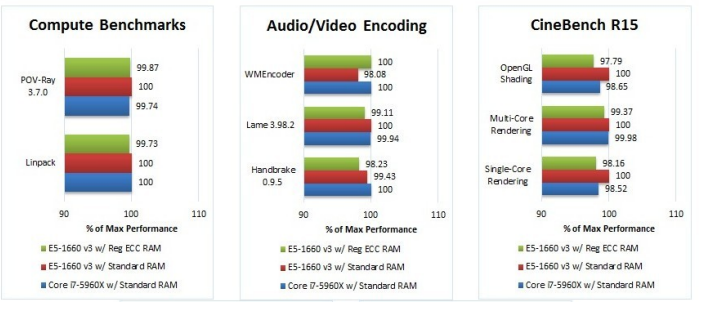




Benchmarking for consumer Adobe CC suite, no point of buying a dual Xeon CPU system, as i7 performs a lot better. (right) Look what happens when the typical work is no longer 100% multi-threaded, but also comprises single-threaded parts, as it does in the real world out there.
http://ppbm7.com/index.php/tweakers-page/95-single-or-dual-cpu/109-single-or-dual-cpu

So why would you ever buy a Core i7-5XXX CPU instead of a Xeon E5 v3?
https://www.pugetsystems.com/labs/articles/Intel-CPUs-Xeon-E5-vs-Core-i7-634/

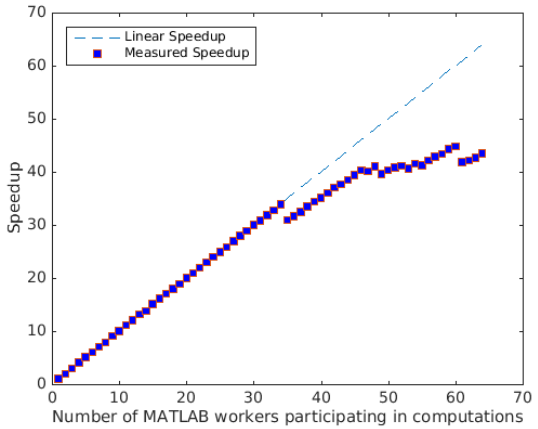
What it comes down to is that Core i7 CPUs are usually slightly cheaper than their Xeon E5 v3 counterparts and they allow for CPU overclocking. If you do not plan on overclocking, we highly recommend you consider **using a Xeon instead of a Core i7 CPU**. You get a much wider range of options - which allows you to get exactly the core count and frequency that is best for your application - and the capability to have huge amounts of system RAM. Even if you don't ever anticipate needing more than 64GB of RAM, having the option for future upgrades is almost never a bad thing.

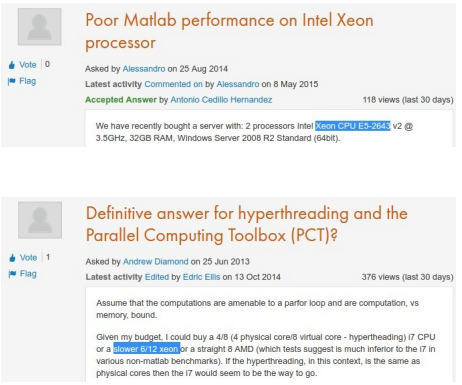


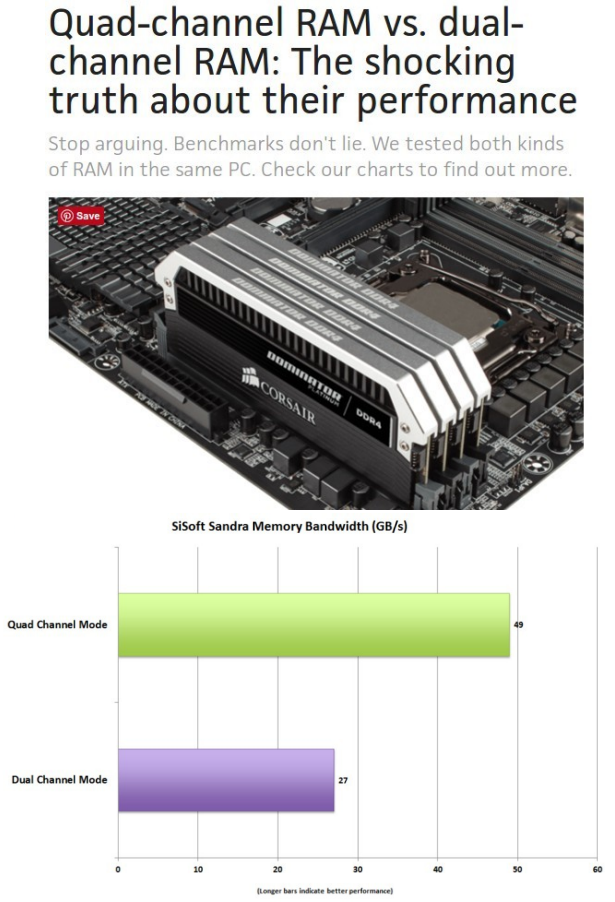
Xeon Performance

- To get the most out of your dual-CPU Xeon system you may have to tweak some MB settings, and remember to populate the DIMM slots in batches of 4 (four-channel DDR4). In other words either 4x or 8x modules **PER CPU.**

http://www.cfd-online.com/Forums/hardware/144936-single-i7-much-faster-than-dual-xeon-e5-2650-v3.html







pcworld.com

CPU Cooler XEON DOES NOT COME WITH ONE

Noctua NH-U9DX i4 Xeon CPU Cooler

NH-U9DXI4 - Noctua NH-U9DX i4 High Performance Intel Xeon CPU Cooler, 2x 92mm, 4U Compatible

Scan Code: **LN64899**

Manufacturer Code: NH-U9DXI4

Scan Rating: High End

Write the first review

Follow this product

Be the first to ask a question.

£39.99 Ex VAT

£47.99 Inc VAT

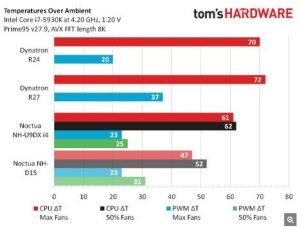
zoom

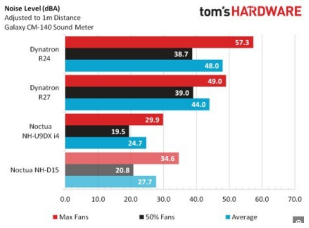
http://noctua.at/en/nh-u9dx-i4

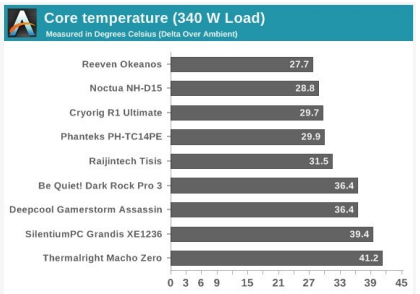
Best Air CPU Coolers

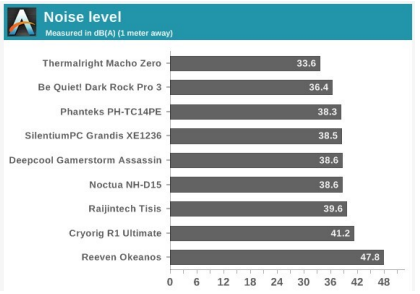
Noctua NH-D14	Be Quiet! Dark Rock Pro 3	Deepcool Gamer Storm Assassin II	Noctua NH-D15	Noctua NH-U14S
Big	BIG	BIG	Big	Big
REVIEW ›	REVIEW ›	REVIEW ›	REVIEW ›	REVIEW ›
£56.90 Amazon	£63.52 Amazon	£79.99 Amazon	£67.98 Amazon	£49.49 Amazon
2012 Editors' Choice	2015 Editor Recommended	2015 Editor Recommended	2014 Editor Recommended	2013 Editor Approved
... more	... more	... more	... more	... more
Height **160mm**	Height **168mm**	Height **167mm**	Height **162mm**	Height **165mm**
Width **130mm**	Width **140mm**	Width **158mm**	Width **152mm**	Width **150mm**
Depth **130mm (158mm w/Fan)**	Depth **122mm (150mm w/Fan)**	Depth **120mm (147mm w/front fan)**	Depth **134mm (162mm w/Fan)**	Depth **52mm (78mm w/Fan)**
Cooling Fans **(1) 120mm x 25mm**	Cooling Fans **120+135mm (25/22mm)**	Cooling Fans **120+140mm x 25mm**	Cooling Fans **(2) 150 x 25mm**	Cooling Fans **(1) 120 x 25mm**

http://www.tomshardware.co.uk/best-cpu-coolers,review-33267.html









NOCTUA NH-U9DX I4

PROS

- The NH-U9DX i4 is capable of cooling overclocked Haswell-E processors at low noise.

CONS

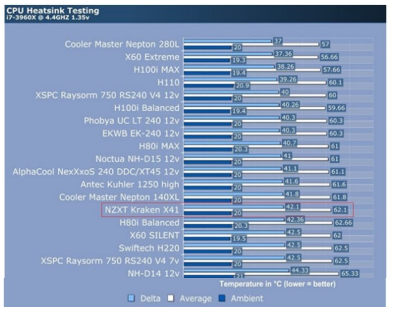
- A maximum DIMM height of less than 35mm and overall cooler height of 5" limit a builder's choice of DRAM and cases

VERDICT

The NH-U9DX i4 provides enough of a performance advantage that we recommend builders to select a case and memory to match. It's because this logic sounds backwards that we gave the NH-U9DX i4 an approved, rather than recommended, award

NH-U9DX i4 work both in 4U Servers and in workstations

tomshardware.co.uk

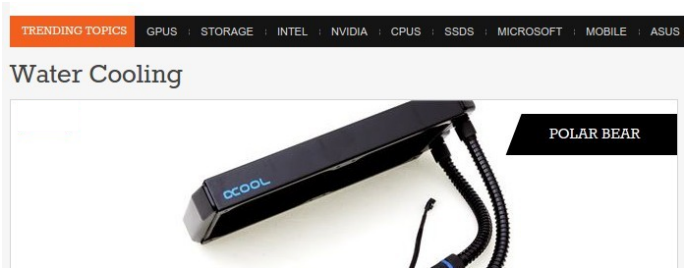


'Torture overclocking' with NZXT Kraken X41 liquid CPU cooler falling behind Noctua NH-D15 for example.

overclock3d.net

Water Cooling

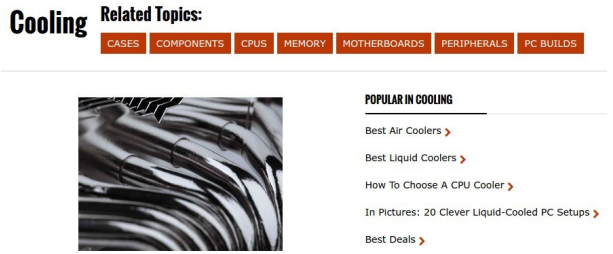
- For better cooling performance of CPU and GPU, you can go for water cooling resulting more silent and cooler system with added cost and harder install



http://www.anandtech.com/tag/water-cooling

Best Closed-Loop Liquid CPU Coolers

tomshardware.com/reviews/best-cpu-coolers

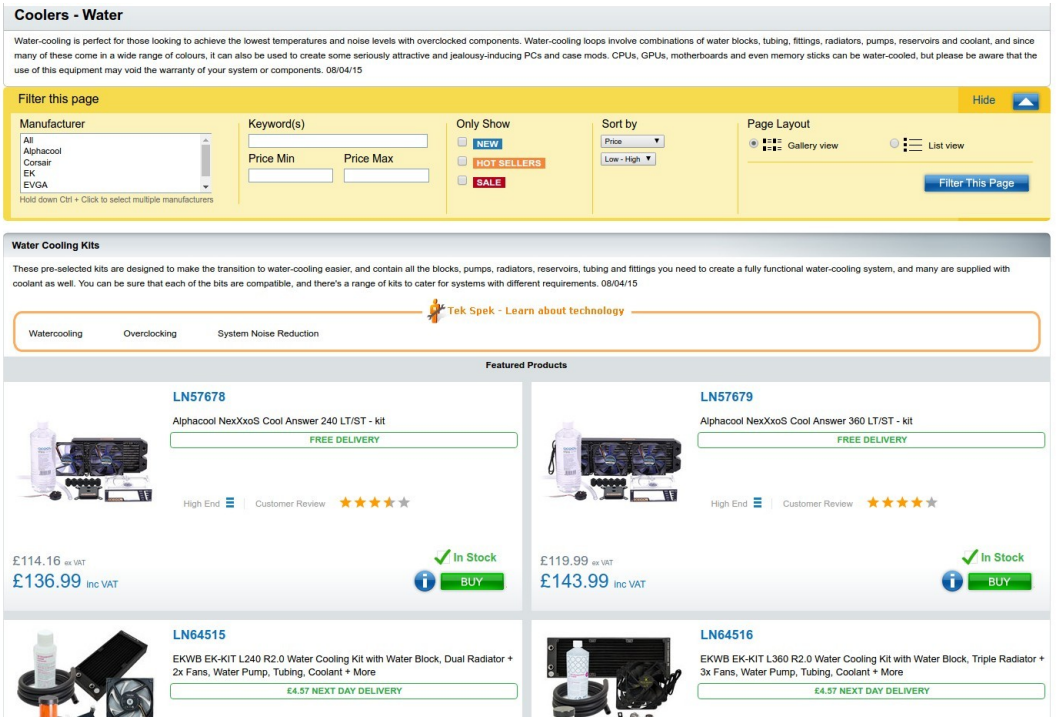


http://www.tomshardware.com/t/cooling/

Roelof Pieters:

"**Cooling is super important**, as it affects both performance and noise. You want to keep the temperature of a GPU at all times below 80 degrees. Anything higher will make the unit lower its voltage and take a hit in performance. Furthermore, too hot temperatures will wear out your GPU; Something you'd probably want to avoid. As for cooling there are two main options: Air cooling (Fans), or Water cooling (pipes):

- **Air cooling** is cheaper, simple to install and maintain, but does make a hell lot of noise;
- **Water cooling** is more expensive, tough to install correctly, but does not make any noise, and cools the components attached to the water cooling system much much better. You would want some chassis fans anyway to keep all the other parts cool, so you'd still have some noise, but less than with a fully air cooled system."



https://www.scan.co.uk/shop/computer-hardware/all/coolers-water/water-cooling-kits

Liquid performance COOLER MASTER NEPTON 280L

The most efficient, but noisy
http://www.frostytech.com/top5_liquid_heatsinks.cfm

Roll over image to zoom in

Cooler Master RL-N28L-20PK-R1 Nepton 280L "Liquid CPU Cooler, Dual JetFlo 140mm fans, Unique LED Illumination on Water Block Lens"
by Cooler Master
15 customer reviews | 9 answered questions

Price: £99.99 & **FREE Delivery** in the UK. Details

Only 8 left in stock - order soon.

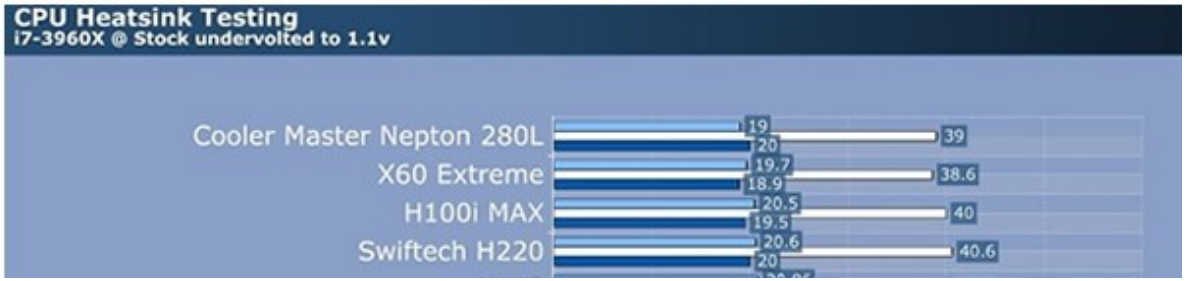
Want it tomorrow, 11 Aug.? Order it within **4 hrs 1 min** and choose **One-Day Delivery** at checkout. Details

Dispatched from and sold by Amazon. Gift-wrap available.

Note: This item is eligible for **click and collect.** Details

2 used from £85.74

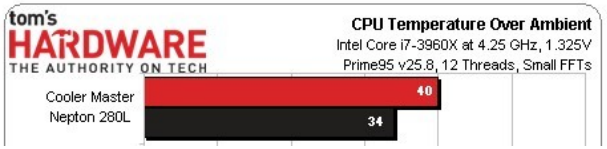
- Factory filled with coolant, then sealed and pressure tested - requires zero maintenance for years
- CM self-designed pump and waterblock guarantees the best water flow and system performance
- Bigger tubing and radiator size provide superior heat dissipation performance
- Exclusive JetFlo 140 fan application to ensure the best radiator heat change efficiency
- Unique LED lens on water block for ID outlook

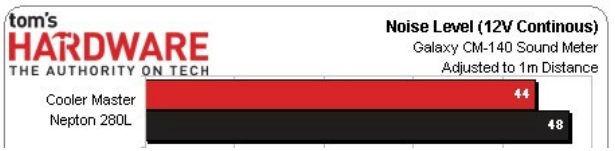


http://www.overclock3d.net/reviews/cases_cooling/nzxt_kraken_x41_review/6

CPU Cooler	Fan Speed	Ambient Temperature	Max CPU Temperature (core average)	Delta Temperature	Noise Level
Cryorig R1 Ultimate	100%	24.50	65.75	41.25	38dB
Noctua NH-D15	100%	24.00	66.50	42.50	44dB
EKWB EK-Kit L240	100%	24.00	66.50	42.50	52dB
Thermalright Archon IB-E X2	100%	25.00	68.00	43.00	35dB
Thermaltake Frio Extreme Silent 14 Dual	100%	26.00	72.50	46.50	33dB
Akasa Medusa Venom	100%	24.00	70.75	46.75	40dB
Gelid The Black Edition	100%	24.50	72.50	48.00	34dB
Cooler Master Nepton 280L	100%	18.00	69.25	51.25	67db
Scythe Tatsumi	100%	24.50	77.25	52.75	32dB
Raijintek Themis	100%	20.00	75.00	55.00	47dB

Intel Core i5-4670K – 4.5GHz (manual overclock via Intel XTU)
pcgameware.co.uk
As soon as you push the i5-4670K above 4.0GHz it certainly starts to warm up, in fact due to the silicon lottery many of these CPUs sadly won't get to the giddy heights of 4.5GHz. Of course those that do, give CPU Coolers the opportunity to show what they can really do. The **Noctua NH-D15 keeps the CPU in check with a maximum average core temperature** of 66.50C (42.50C Delta). Which again shows the NH-D15 can handle and tame a hot CPU with ease.





tomshardware.co.uk

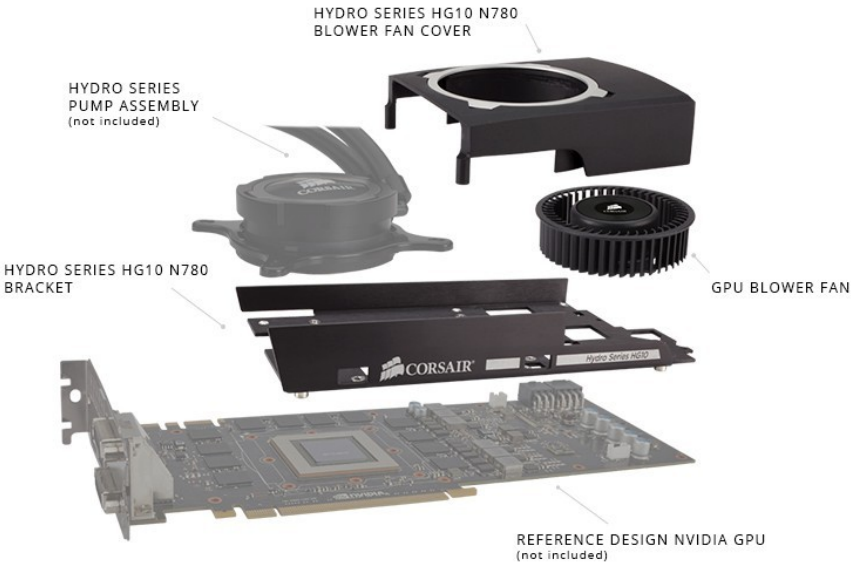
Not really worth the money for Intel Xeon (85W each) that won't be intensively overclocked nor overused really.

Go for Noctua NH-D15 or U9DX

GPU Liquid Cooling OEM NVIDIA SOLUTIONS

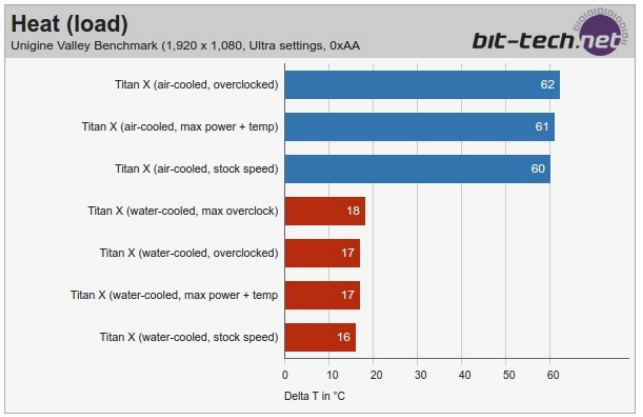
evga.com

EVGA GeForce GTX TITAN X HYBRID



Hydro Series HG10 N780 GPU Liquid Cooling Bracket

corsair.com



bit-tech.net

Water-cooling Nvidia's Titan X - Testing, Cooling and Performance.

Published on 4th May 2015 by Antony Leather

GPU Cooling ECONOMICS OF COOLING

- The sub-20°C GPU temperatures obtained with water cooling are an epic achievement compared to over 60°C with conventional air cooling (previous slide, bit-tech.net)
  - Higher temperature, and temperature cycling in general age the electronics components and lead to earlier failures.
  - There is literature on *prognostics and health management* assigning cost to failing electronics. Especially important in data centers, military electronics, etc. that need high reliability.
  - Personally I don't know any study trying to quantify the reasonable investment cost of liquid cooling in terms of lice cycle assessment.

Nvidia's GPU failures: A case for prognostics and health management

Michael Pecht
Prognostics and Systems Health Management Center, City University of Hong Kong, Hong Kong
CALCE Electronics Products and Systems Center, University of Maryland, College Park, MD 20742, USA

http://dx.doi.org/10.1016/j.microrel.2011.11.017

Prognostics and health management of electronics

NM Vichare, MG Pecht - *IEEE transactions on components and packaging ...*, 2006 Abstract—There has been a growing interest in monitoring the ongoing "health" of products and systems in order to predict failures and provide warning to avoid catastrophic failure. Here, health is defined as the extent of degradation or deviation from an expected normal ... Cited by 348 Related articles

Benefits and challenges of system prognostics

B Sun, S Zeng, R Kang, MG Pecht - *IEEE Transactions on reliability*, 2012 Abstract—Prognostics is an engineering discipline utilizing in-situ monitoring and analysis to assess system degradation trends, and determine remaining useful life. This paper discusses the benefits of prognostics in terms of system life-cycle processes, such as ...Cited by 55 Related articles

Prognostics health management: perspectives in engineering systems reliability prognostics

MAM Esteves, EP Nunes - *Safety and Reliability of Complex Engineered Systems*, 2015The Prognostic Health Management (PHM) has been asserting itself as the most promising methodology to enhance the effective reliability and availability of a product or system during its life-cycle conditions by detecting current and approaching failures, thus, providing ... Related articles

Google Scholar

Predicting overtemperature events in graphics cards using regression models

FCM Rodrigues, LP Queiroz, JC Machado - 2015 Brazilian Conference on ..., 2015 Abstract—Graphics cards are complex electronic systems designed for high performance applications. Due to its processing power, graphics cards may operate at high temperatures, leading its components to a significant degradation level. This fact is even more present ...Related articles

In situ temperature measurement of a notebook computer-a case study in health and usage monitoring of electronics

N Vichare, P Rodgers, V Eveloy, MG Pecht - IEEE Transactions on Device and ..., 2004 Abstract—Reliability prediction methods do not generally ac-count for the actual life cycle environment of electronic products, which covers their environmental, operating and usage conditions. Considering thermal loads, thermal management strategies still focus on a ...Cited by 83 Related articles

Baseline performance of notebook computers under various environmental and usage conditions for prognostics

S Kumar, M Pecht - IEEE Transactions on Components and Packaging ..., 2009Abstract—This paper presents an approach for electronic product characterization. A study was conducted to formally characterize notebook computer performance under various environmental and usage conditions. An experiment was conducted on a set of ten ...Cited by 12 Related articles

Health-monitoring method of note PC for cooling performance degradation and load assessment

K Hirohata, K Hisano, M Mukai - Microelectronics Reliability, 2011 Health monitoring technologies, which can evaluate the performance degradation, load history and degree of fatigue, have the potential to improve the effective maintenance, the reliability design method and the availability in the improper use conditions of electronic ...Cited by 4 Related articles

A circuit-centric approach to electronic system-level diagnostics and prognostics

ASS Vasan, C Chen, M Pecht - Prognostics and Health Management (PHM), 2013 ..., 2013Abstract—Electronic system failures during field operation in mission, safety and infrastructure critical applications can have severe implications. In these applications, incorporating prognostics and health management (PHM) techniques provide systems ...Cited by 1 Related articles

Limits of air-cooling: status and challenges

P Rodgers, V Eveloy, MG Pecht - ... Measurement and Management IEEE Twenty First ..., 2005Abstract Despite the perception that the limits of air-cooling have been reached, this paper reviews approaches that could maintain the effectiveness of this technology. Key thermal management areas that need to be addressed are discussed, including heat sink design ...Cited by 28 Related articles

Enabling electronic prognostics using thermal data

N Vchare, M Pecht - arXiv preprint arXiv:0709.1813, 2007Abstract: Prognostics is a process of assessing the extent of deviation or degradation of a product from its expected normal operating condition, and then, based on continuous monitoring, predicting the future reliability of the product. By being able to determine when ... Cited by 23 Related articles

Real-time temperature estimation for power MOSFETs considering thermal aging effects

H Chen, B Ji, V Pickert, W Cao - IEEE Transactions on Device and Materials ..., 2014 Abstract—This paper presents a novel real-time power device temperature estimation method which monitors the power MOSFET's junction temperature shift arising from thermal aging effects and incorporates the updated electrothermal models of power modules into ...Cited by 20 Related articles

The benefits of data center temperature monitoring

MRC Truşcă, Ş Albert, ML Soran - Grid, Cloud & High Performance Computing in ..., 2015 Abstract—The temperature is an important parameter for the equipment functioning. Computer systems are designed to work best when the ambient temperature is in the range 20–23 C. This requirement is ensured by automated systems that maintain the ...Related articles

Health assessment of electronic products using Mahalanobis distance and projection pursuit analysis

S Kumar, V Sotiris, M Pecht - International Journal of Computer, Information, and ..., 2008 Abstract—With increasing complexity in electronic systems there is a need for system level anomaly detection and fault isolation. Anomaly detection based on vector similarity to a training set is used in this paper through two approaches, one the preserves the original ...Cited by 20 Related articles

CPU ACCELERATION FOR HIGH MEMORY NEEDS

Using Intel's Xeon Phi for Brain Research Visualization

Ron Farber, July 11, 2016, 7:55 a.m.

Swiss Blue Brain Project at the École Polytechnique Fédérale de Lausanne (EPFL).
www.top500.org | news.ycombinator.com

The Blue Brain demo was one of five live, interactive visualization demonstrations running on an Intel Scalable Systems Framework cluster using the new Intel 7210 Xeon Phi ("Knights Landing") processors.

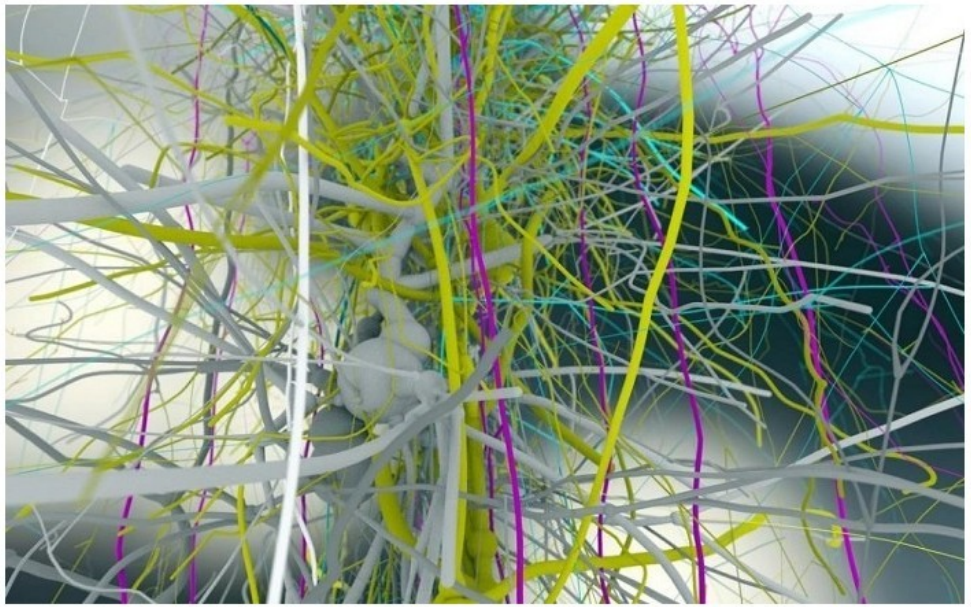


Figure 1: Even first in-silico models show the complexity and beauty of the brain

Ray-tracing provides a natural way to work with these parametric descriptions so the image can be efficiently rendered in a *much smaller amount of memory*. However, 'smaller' is a relative term as current visualizations can occur on a machine that contains less than a terabyte of RAM. Traditional *raster-based* rendering would have greatly increased the memory consumption as the convoluted shape of each neuron would require a mesh containing approximately 100,000 triangles per neuron. Scientists currently visualize biological networks containing 1,000 to 10,000 neurons.

IBM to deliver 200-petaflop supercomputer by early 2018; Cray moves to Intel Xeon Phi

By Jamie Lendino on June 22, 2016 at 9:47 am | 82 Comments

3.2K shares

http://www.extremetech.com/

Forbes / Tech

JUN 28, 2016 @ 11:00 AM 6,970 VIEWS

The Little Black Book of Billionaire Secrets

Can Intel's New Knights Landing Chip Compete With NVIDIA For Deep Learning?

Moor Insights and Strategy, CONTRIBUTOR

Straight talk from Moor Insights & Strategy tech industry analysts

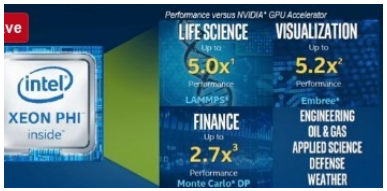
FULL BIO

Opinions expressed by Forbes Contributors are their own.

POST WRITTEN BY

Karl Freund

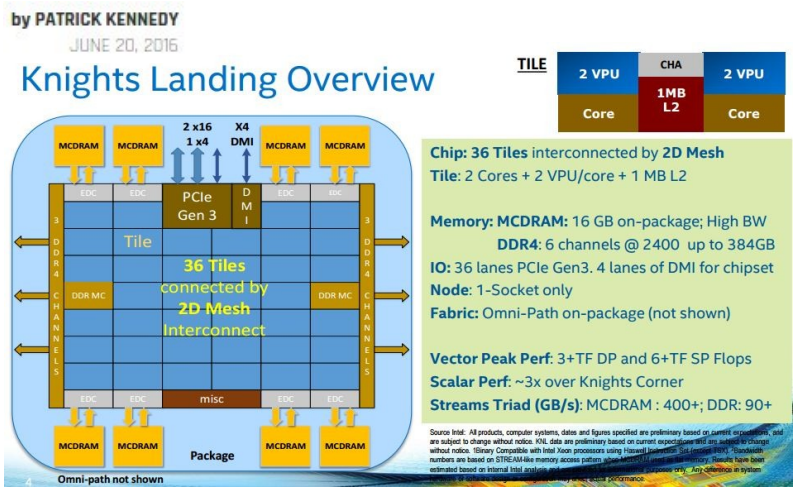
Karl Freund is a Moor Insights & Strategy analyst for machine learning & HPC

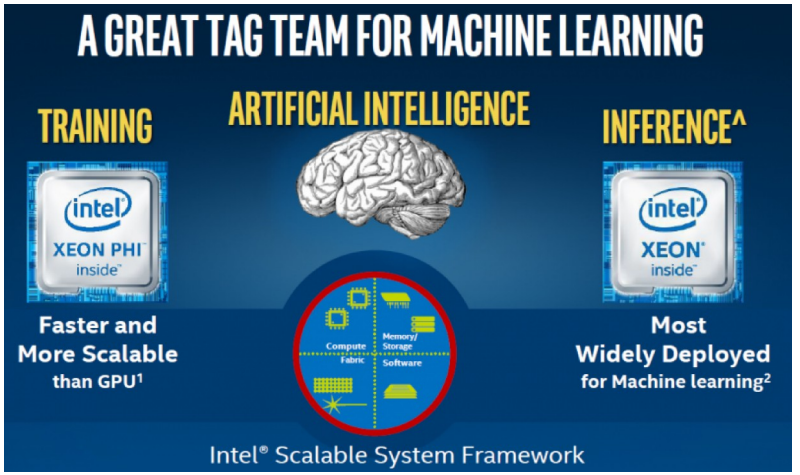


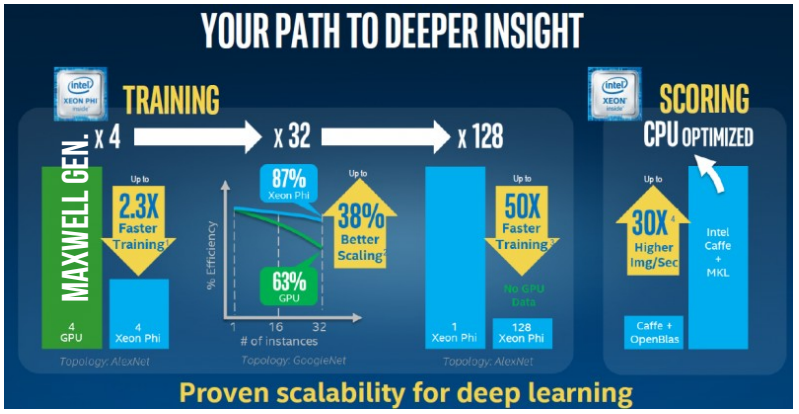
forbes.com

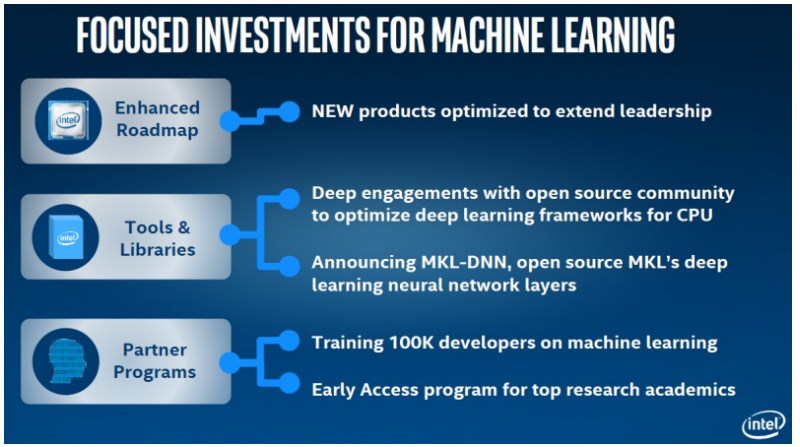
CPU Machine Learning?

Intel Strikes Back with Intel Xeon Knights Landing









http://www.servethehome.com/intel-xeon-phi-knights-landing-generation-now-available/

Getting Started with the Intel Xeon Phi on Ubuntu 14.04/Linux Kernel 3.13.0

Intel Xeon Phi on Ubuntu 14.04 (not officially supported)
http://arrayfire.com/getting-started-with-the-intel-xeon-phi-on-ubuntu-14-04linux-kernel-3-13-0/

PETER / DECEMBER 11, 2014 / HARDWARE, OPENCL / 19 COMMENTS

CPU Machine Learning Benchmarks

NVIDIA responds

Correcting Intel's Deep Learning Benchmark Mistakes

Posted on AUGUST 16, 2016 by IAN BUCK

Benchmarks are an important tool for measuring performance, but in a rapidly evolving field it can be difficult to keep up with the state of the art. Recently Intel published some incorrect "facts" about their long promised Xeon Phi processors.

https://blogs.nvidia.com/blog/2016/08/16/correcting-some-mistakes/

For example, Intel recently published some out-of-date benchmarks to make three claims about deep learning performance with Knights Landing Xeon Phi processors:

1) Xeon Phi is 2.3x faster in training than GPUs(1)

2) Xeon Phi offers 38% better scaling that GPUs across nodes(2)

3) Xeon Phi delivers strong scaling to 128 nodes while GPUs do not(3)

We'd like to address these claims and correct some misperceptions that may arise.

Intel used Caffe AlexNet data that is 18 months old, comparing a system with four Maxwell GPUs to four Xeon Phi servers. With the more recent implementation of Caffe AlexNet, publicly available here, Intel would have discovered that the same system with four Maxwell GPUs delivers 30% faster training time than four Xeon Phi servers.

In fact, a system with four Pascal-based NVIDIA TITAN X GPUs trains 90% faster and a single NVIDIA DGX-1 is over 5x faster than four Xeon Phi servers.

System	Hours to Train	Speed-up vs Xeon Phi
PC with 4x NVIDIA TITAN X (Maxwell) ** based on NVIDIA Caffe implementation as of March 2015*	25 Hours	-
Four Xeon Phi servers	10.5 Hours	-
PC with 4 NVIDIA TITAN X (Maxwell) ** based on publicly available Caffe as of August 2016, cuDNN5*	8.2 Hours	1.3x faster than Xeon Phi
PC with 4 NVIDIA TITAN X (Pascal) ** based on publicly available Caffe as of August 2016, cuDNN5*	5.5 Hours	1.9x faster than Xeon Phi
NVIDIA DGX-1	2 Hours	5.3x faster than Xeon Phi

FPGAs and ASICs in Deep Learning Short intro

CNNLab: a Novel Parallel Framework for Neural Networks using GPU and FPGA

— a Practical Study with Trade-off Analysis

Maohua Zhu, Liu Liu
Electrical and Computer Engineering, UCSB
Email:{maohuazhu,liu_liu}@umail.ucsb.edu

Chao Wang
Computer Sciecne,USTC
Email:cswang@ustc.edu.cn

Yuan Xie
Electrical and Computer Engineering,UCSB
Email:yuanxie@ece.ucsb.edu

http://arxiv.org/abs/1606.06234

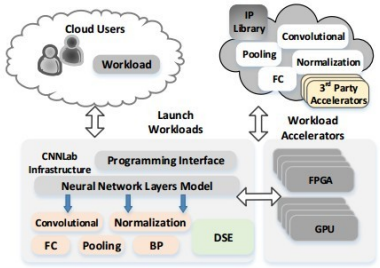


Fig. 2. High level CNNLab architecture with corresponding neural network layers model. The NN processing are decomposed into layers and scheduled either on the GPU or FPGA-based accelerators.

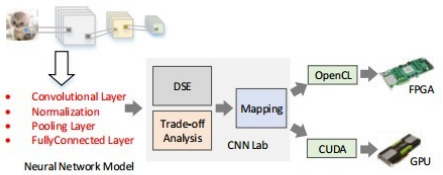
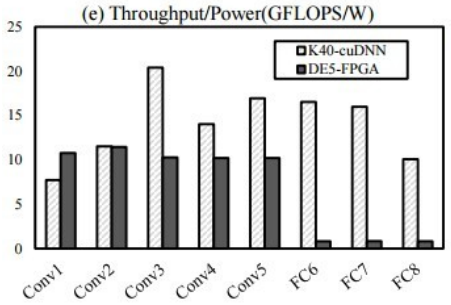
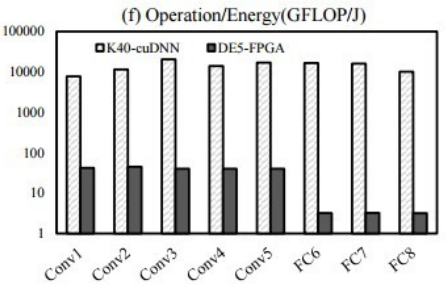


Fig. 3. General Processing Flow of the CNNLab.





EMERGING "UNIVERSAL" FPGA, GPU PLATFORM FOR DEEP LEARNING

June 29, 2016 Nicole Hemsoth

http://www.nextplatform.com/2016/06/29

Computer Science > Distributed, Parallel, and Cluster Computing

Deep Learning on FPGAs: Past, Present, and Future

Griffin Lacey, Graham W. Taylor, Shawki Areibi

(Submitted on 13 Feb 2016)

http://arxiv.org/abs/1602.04283

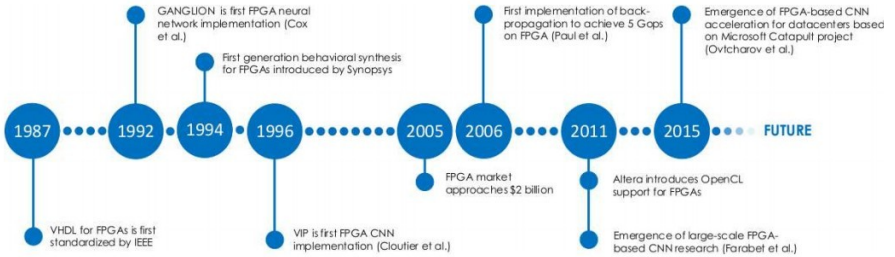


Figure 3: Timeline of important events in FPGA deep learning research.

Table 1: Overview of Deep Learning Frameworks with OpenCL Support

Tool	Core Language	Bindings	OpenCL	User Base
Caffe	C++	Python MATLAB	Partial Support	Large
Torch	Lua	-	Partial Support	Large
Theano	Python	-	Minimal Support	Large
DeepCL	C++	Python Lua	Full Support	Moderate

ASICs for DNNs

- Exponential growth in Data Centers
  - Commoditization of Enterprise silicon
  - Traditional mobile players announcing ASICs for enterprise compute
- Higher demands for compute density
  - GPGPUs have won the first round
  - Dennardian scaling is breaking down
- Power dissipation will emerge as major challenge
  - Chip level, server level and DC level

Sumit Sanyal CEO, minds.ai
ip.cadence.com

Xeon Phi Deep Learning

Xeon-CafPhi

Caffe deep learning framework - optimized for Xeon Phi

By: Rohith Jagannathan, Dhruv Saksena

How Intel® Xeon Phi™ Processors Benefit Machine Learning/Deep Learning Apps and Frameworks

By Pradeep Dubey (Intel), June 20, 2016 Translate

https://software.intel.com/en-us/blogs/2016/06/20

arXiv.org > cs > arXiv:1510.06706 Search or Artic

Computer Science > Neural and Evolutionary Computing

ZNN - A Fast and Scalable Algorithm for Training 3D Convolutional Networks on Multi-Core and Many-Core Shared Memory Machines

Aleksandar Zlateski, Kisuk Lee, H. Sebastian Seung

(Submitted on 22 Oct 2015)

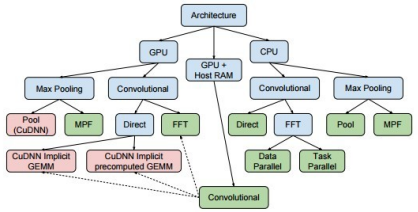


Fig. 1: Diagram of all layers primitives. The red primitives are wrappers primitives provided by CuDNN. The green primitives are the novel primitives introduced in this paper.

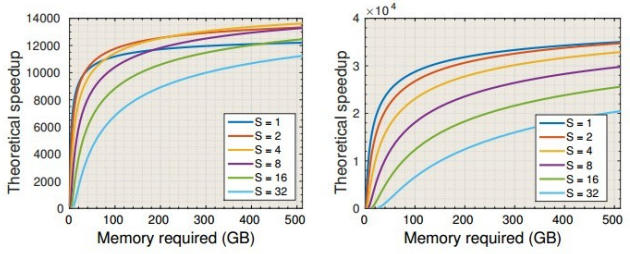


Fig. 4: Theoretical speedup of pooling networks using FFT–based convolution for different sizes of the inputs and different batch sizes for a network with 1 pooling layer (a) and 2 pooling layers (b).

arXiv.org > cs > arXiv:1606.05688 Search or Artic

Computer Science > Distributed, Parallel, and Cluster Computing

ZNNi - Maximizing the Inference Throughput of 3D Convolutional Networks on Multi-Core CPUs and GPUs

Aleksandar Zlateski, Kisuk Lee, H. Sebastian Seung

(Submitted on 17 Jun 2016)

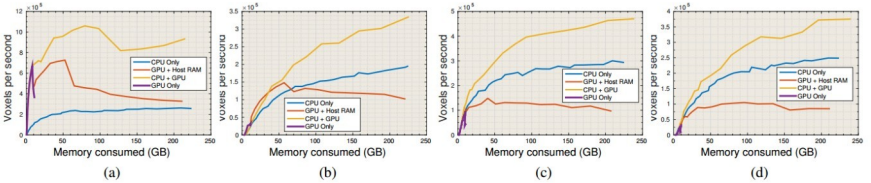


Fig. 7: Maximal throughput achieved vs memory consumption using GPU–only, CPU–only, CPU + host RAM and CPU–GPU implementations for different image sizes.

Network	Baseline (cuDNN)	Caffe	ELEKTRONN	ZNN	GPU-Only	CPU-Only	GPU + host RAM	GPU-CPU
n337	22,934.8	1.348	122,668	34,334.8	671,782	262,131	727,103	1,059,910
n537	1,048.68	–	–	9,494.5	29,352.1	194,683	147,965	334,163
n726	13,520.4	–	6,122	31,354.8	97,257.2	300,312	148,194	470,166
n926	2,667.86	–	–	20,908.6	35,051.3	249,190	104,946	375,295

TABLE V: Comparisons to other methods.

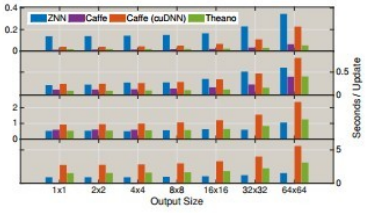


Fig. 8: Comparison of ZNN, Caffe (with and without cuDNN) and Theano for 2D ConvNets. The charts from the top down have kernel sizes of 10^2, 20^2, 30^2 and 40^2 respectively. Where Caffe data is missing, it means that Caffe could not handle networks of the given size.

arXiv.org > cs > arXiv:1508.04843 Search or Artic

Computer Science > Computer Vision and Pattern Recognition

Recursive Training of 2D-3D Convolutional Networks for Neuronal Boundary Detection

Kisuk Lee, Aleksandar Zlateski, Ashwin Vishwanathan, H. Sebastian Seung

(Submitted on 20 Aug 2015)

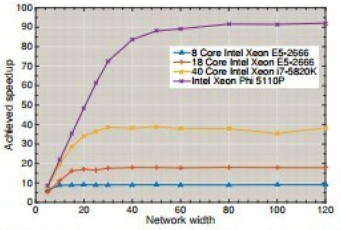


Fig. 7: Achieved speedups on 3D networks compared to the serial algorithm.

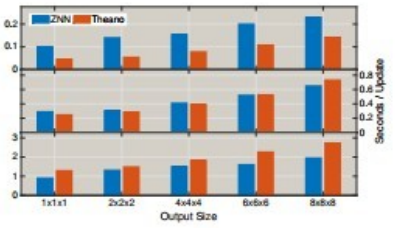


Fig. 9: Comparison of ZNN and Theano for 3D ConvNets. The charts from the top down have kernel sizes of 3^3, 5^3 and 7^3.

ZNN

(Seung lab, Lee, Zlateski, Wu, Turner)

https://github.com/seung-lab/znn-release

Excels with large kernel sizes when compared to other frameworks. Especially useful with 3D datasets such as electron microscopy, MRI, etc.

Intel acquires ML hardware #1

NERVANA with ASIC chip coming up

MIT Technology Review

Topics+ Top Stories

Business

Intel Buys a Startup to Catch Up in Deep Learning

Acquisition should let Nervana Systems speed development of its chips radically redesigned for artificial intelligence.

by Peter Burrows and Tom Simonite August 9, 2016

technologyreview.com

Intel acquires artificial intelligence start-up in tech expansion

Purchase of Nervana will help chipmaker extend its reach into software and chips

ft.com

TECH INTERNET OF THINGS See the Fortune 500 list

Why Intel Bought Artificial Intelligence Startup Nervana Systems

by Aaron Pressman @ampressman AUGUST 9, 2016, 4:00 PM EDT

fortune.com

ALTERA with FPGA chips

TECH POINTCLOUD

Official At Last: Intel Completes $16.7 Billion Buy of Altera

by Barb Darrow @gigabarb DECEMBER 28, 2015, 1:33 PM EDT

http://fortune.com/2015/12/28/intel-completes-altera-acquisition/

Home / Hardware

The first fruits of Intel's huge $16.7 billion Altera buy will come this quarter

The first chips that combine Intel and Altera components will ship in the next few months.

Agam Shah Jan 14, 2016 3:55 PM pcworld.com
IDG News Service

GPU Computation Field Programmable Gate Arrays (FPGAs) +5

Is implementing deep learning on FPGAs a natural next step after the success with GPUs?

Answer Request Follow 98 Comment 1 Share 8 Downvote

13 Answers

Roman Trusov, Facebook AI Research Intern 2016
6.6k Views · Most Viewed Writer in Machine Learning with 210+ answers

It's a big step in preformance, due to amazing computational results that FPGAs can provide. And it's a great engineering advance for deep learning field.
But I don't think there is something fundamentally new here, since it's the same algorithm implemented in an extremely low-level language. There's no new scientific result here, despite the NNs themselves can work better.

As a real "next step" I would rather suggest Jeffrey Dean's result in distributed deep learning: Page on googleusercontent.com

Written Dec 2, 2014 · View Upvotes

Upvote 4 Downvote Comment

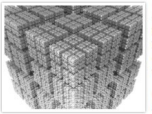
quora.com

THENEXTPLATFORM

HOME COMPUTE STORE CONNECT CONTROL CODE ANALYZE HPC ENTERPRISE

EMERGING "UNIVERSAL" FPGA, GPU PLATFORM FOR DEEP LEARNING

June 29, 2016 Nicole Hemsoth



In the last couple of years, we have written and heard about the usefulness of GPUs for deep learning training as well as, to a lesser extent, custom ASICs and FPGAs. All of these options have shown performance or efficiency advantages over commodity CPU-only approaches, but programming for all of these is often a challenge.

Programmability hurdles aside, deep learning training on accelerators is standard, but is often limited to a single choice—GPUs or, to a far lesser extent, FPGAs. Now, a research team from the University of California Santa Barbara has proposed a new middleware platform that can combine both of those accelerators under a common programming environment that creates enough abstraction over both devices to allow a convolutional neural network to leverage both with purported ease.

nextplatform.com

Intel acquires ML hardware #2

Forbes

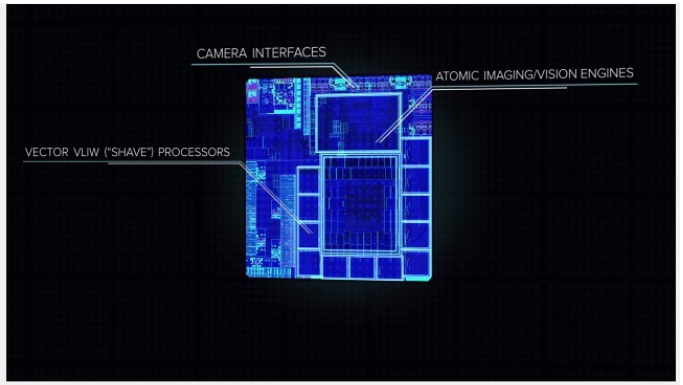
SEP 8, 2016 @ 01:40 PM 994 VIEWS The Little Black Book of Billionaire Secrets

Intel Boosts Its Perceptual Computing Strategy With The Movidius Acquisition

http://www.forbes.com/sites/greatspeculations/2016/09/08

Earlier this week, leading semiconductor maker, Intel announced the acquisition of Movidius, a company that specialises in designing low power chips for computer vision. Intel already has its own RealSense cameras that feature the groundbreaking depth-sensing technology, which allows devices to "see" the world in three dimensions. With the acquisition of Movidius, the company should be able to take its computer vision and perceptual computing strategy to the next level.

intel + Movidius



Movidius and Google Bring Machine Intelligence to Devices

Movidius

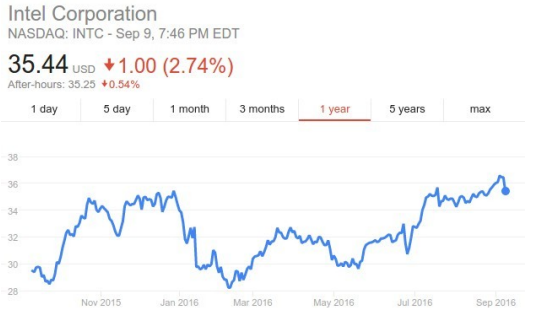
Subscribed 535

52,945 views

Add to Share More 199 10

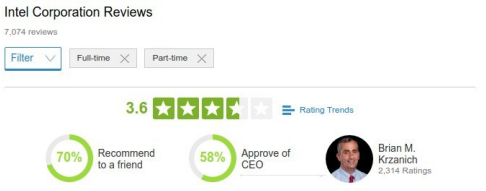
https://www.youtube.com/watch?v=GEy-vtev1Bw

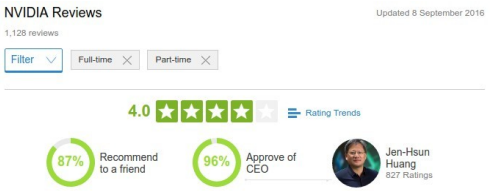
intel NVIDIA



Market cap: $167.57B

Market cap: $32.07B





Intel	$10.1 billion per year (US dollars per year) (trailing 12-month value as of June 30, 2016)
NVIDIA	$903 million per year (US dollars per year) (trailing 12-month value as of July 31, 2016)
Advanced Micro Devices Inc	– $339 million per year (US dollars per year) (trailing 12-month value as of June 30, 2016)

http://www.wolframalpha.com/input/?i=intel+earnings+vs+nvidia+earnings+vs+amd+earnings

Intel Future releases

INTEL SETS UP SKYLAKE XEON FOR HPC, KNIGHTS MILL XEON PHI FOR AI

November 15, 2016 Timothy Prickett Morgan

https://www.nextplatform.com/2016/11/15/intel-sets-skylake-xeon-hpc-knights-mill-xeon-phi-ai/

Intel® Deep Learning Inference Accelerator

Integrated hardware and software solution to accelerate convolutional neural networks

- Simplify deployment with preloaded, optimized algorithms
- PCIe card with Intel® Arria® 10 FPGA
- Software programmable through standard frameworks and libraries
- Coming in 2017

Knights Mill and an FPGA adapter card that Intel has created and paired with FPGA Verilog "code" that handles machine learning inference routines. The idea is to use Knights Mill for training neural nets and this FPGA card and its "software" to handle machine learning inference offload for applications, which Intel will update like other pieces of enterprise software and which Davis said offered four times the performance per watt on inference than a CPU alternative. We will presumably hear more about the Broadwell Xeon-Arria 10 hybrid chip that has been in the works for a while now.

Next Generation Intel® Xeon® Processor

General Availability in mid-2017

HPC Optimizations:

Intel® Advanced Vector Instructions-512 boost floating point calculations & encryption algorithms

Integrated Intel® Omni-Path Architecture for high speed network

The future "Skylake" Xeon E5 v5 processors due to be announced in the middle of next year sometime, will support the 512-bit Advanced Vector Instructions (AVX-512) floating point and encryption features that made their debut first in the Knights Landing chips. As we told you previously, the Skylake Xeon processors are expected to have a single socket for machines that span from two to eight sockets, so in a way, there is no Xeon E5 separate from the Xeon E7 anymore

Next Gen Intel® Xeon Phi™ Processor

Codenamed "Knights Mill"

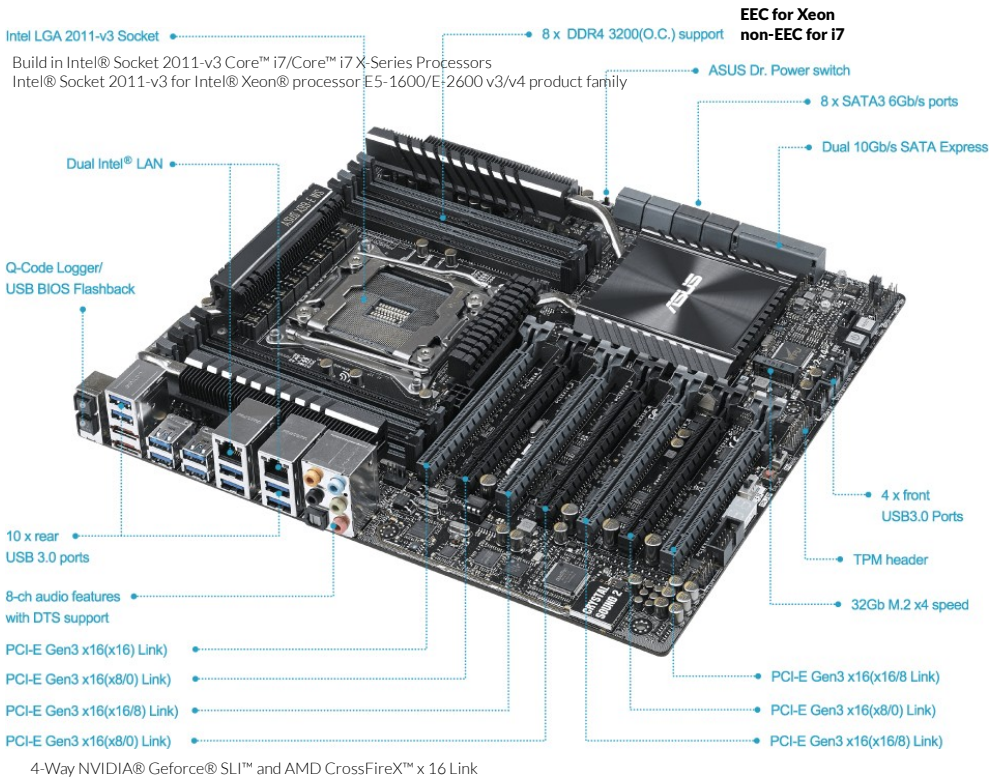
Optimized for Artificial Intelligence

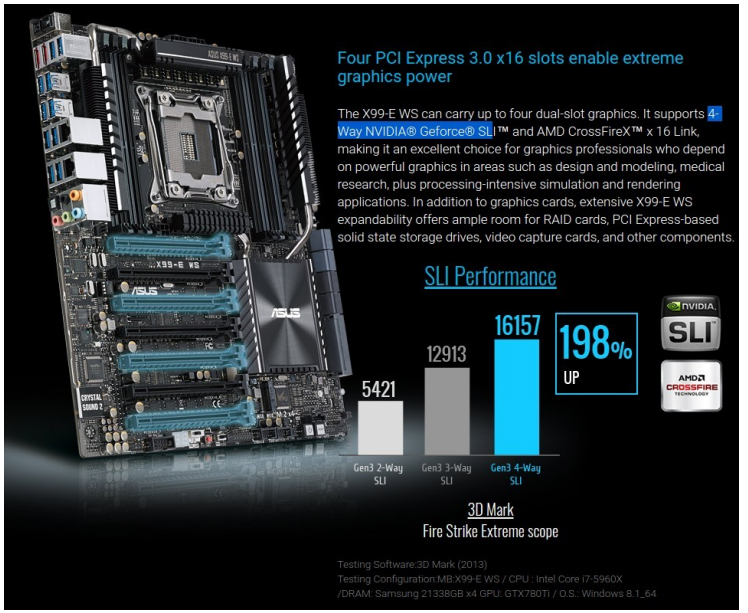
Host-CPU with mixed precision performance for improved machine learning

Coming in 2017

Motherboard Single CPU I7/XEON E5-16XX

Asus X99-E WS, asus.com/uk





NVIDIA® DIGITS™ DevBox https://developer.nvidia.com/devbox

DIGITS DevBox includes:

- Four TITAN X GPUs with 12GB of memory per GPU
- 64GB DDR4
- Asus X99-E WS workstation class motherboard with 4-way PCI-E Gen3 x16 support
- Core i7-5930K 6 Core 3.5GHz desktop processor
- Three 3TB SATA 6Gb 3.5" Enterprise Hard Drive in RAID5
- 512GB PCI-E M.2 SSD cache for RAID
- 250GB SATA 6Gb Internal SSD
- 1600W Power Supply Unit from premium suppliers including EVGA
- Ubuntu 14.04

NVIDIA ACCELERATED COMPUTING Downloads Training

Home › CUDA ZONE › Forums › Accelerated Computing › CUDA Setup and Installation › View Topic

motherboard recommendation for multi-gpu setup

3XS Deep Learning G10 includes:

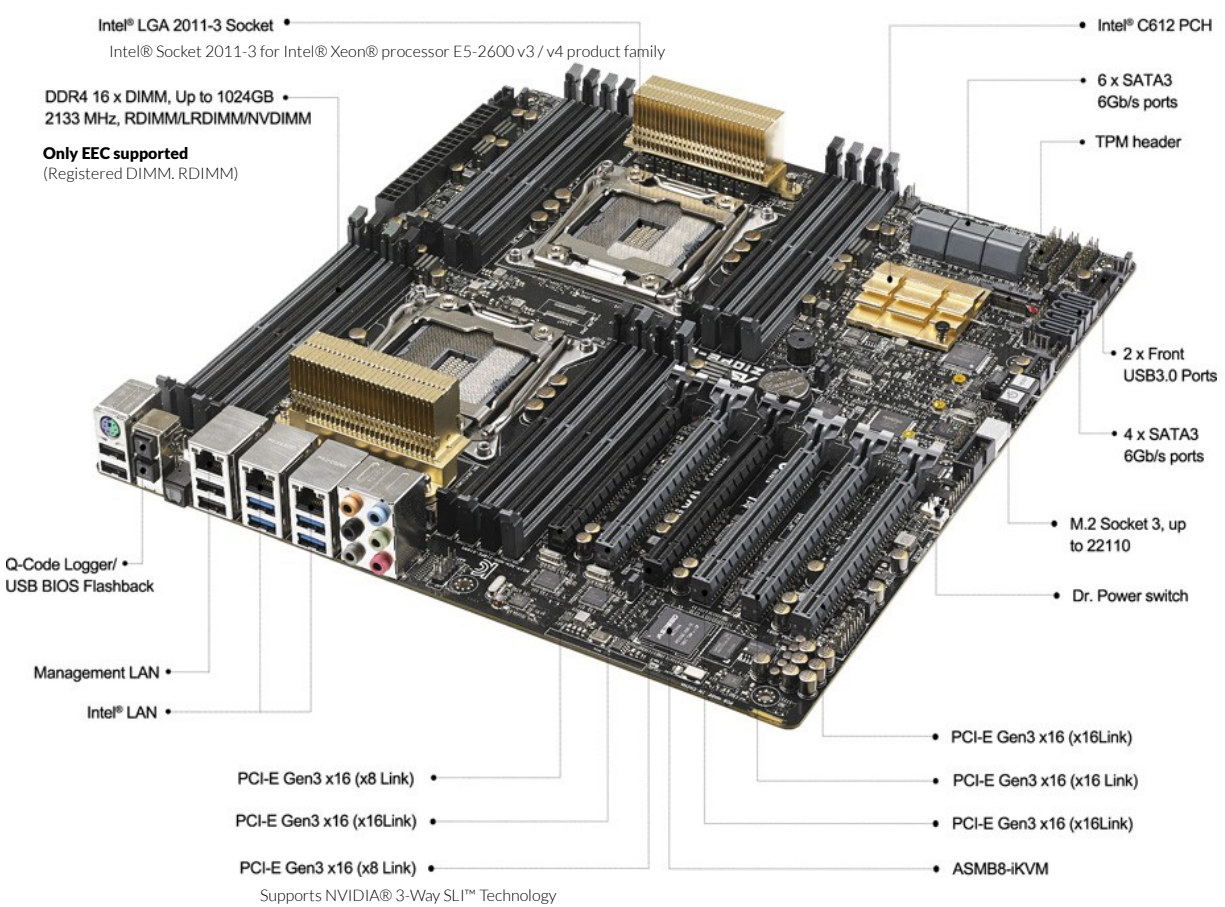
- Four TITAN X GPUs with 12GB of memory per GPU
- 64GB DDR4
- Asus X99-E WS workstation motherboard
- i7 5960X 8-Core 3GHz desktop processor
- Three 3TB SATA 6Gb 3.5" Enterprise Hard Drives
- 512GB PCI-E M.2 SSD cache for RAID
- 250GB SATA 6Gb Internal SSD
- 1600W Power Supply Unit from premium suppliers
- Ubuntu 14.04
- NVIDIA-qualified driver
- NVIDIA® CUDA® Toolkit 7.0
- NVIDIA® DIGITS™ SW
- Caffe, Theano, Torch, BIDMach

CONFIGURE

https://www.scan.co.uk/shops/nvidia/deep-learning

Motherboard Dual XEON CPU ASUS "CONSUMER"

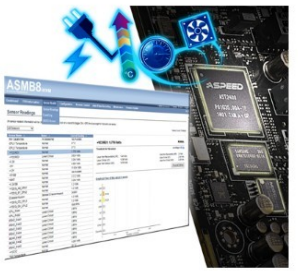
Z10PE-D16 WS, asus.com/uk



only 6 PCIe slots → 3 GPUs

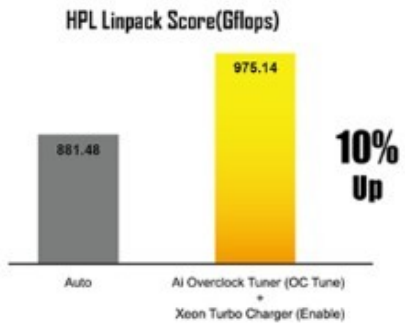
Rich expansion capability with 6 PCI-E Gen3 x16

Six PCI-E Gen3 x16 Slots gives you the sufficient I/O interfaces to fulfill your demand for graphic or computing solution. You'll be able to run both multi-GPU setups. The board features SLI on demand technology, not only supporting up to three graphics cards in a 3-Way SLI but also supporting up to Quad GPU CrossfireX™ technology.



Complete remote server management

The IPMI 2.0-compliant ASMB8-iKVM module enables remote BIOS updates, standalone KVM/Java utility, video recording and BSOD capture. With out-of-band management ASMB8-iKVM provides round-the-clock remote monitoring and diagnosis even if the server operating system is down or offline, and all through a user-friendly web-based graphical interface that works with all major browsers. In addition, ASWM Enterprise software provides one-to-multiple centralized management including BIOS flash, software dispatch, task scheduler, remote control and power control, via colorful and informative graphical interface.



Dual CPU overclocking for incredible performance

Z10PE-D16 WS breaks performance barriers by giving you the capability to overclock dual CPUs. Ai Overclock Tuner and Xeon Turbo Charger in the BIOS are dedicated to boosting the CPU's overclocking performance by up to 10% to reach incredible CPU High Performance Linpack (HPL) scores.

Configuration:
HPL Linkpack Test Configuration:
MB:Z10PE-D16 WS / CPU : Intel Xeon E5-2699 V3/DRAM: Samsung 8GB DDR4-2133 RDIMM*16 /O.S: Windows® Server 2012 R2

CPU, Chipset and Graphics features

Intel® Xeon® E5-2600 processor family for the LGA 2011 socket

The motherboard supports the latest Intel® Xeon® processor E5-2600 v3 product family with dual LGA 2011-3 sockets. Memory and PCI Express controllers are integrated alongside quad-channel 16-DIMM DDR4 memory and 80 PCI Express 3.0 lanes. This provides great graphics performance with superior energy efficiency.

Intel® C612 Express Chipset

The Intel® C612 Chipset is the latest single-chipset design that supports Intel® Xeon® processor E5-2600 v3 product family for dual LGA 2011-3 sockets. It improves performance by utilizing serial point-to-point links, allowing for increased bandwidth and stability. Additionally, the C612 comes with 10 x SATA 6Gbit/s ports for faster data retrieval, doubling the bandwidth of current bus systems.

Supports SLI™ and CrossFireX™

Both SLI™ and CrossfireX™ architectures work flawlessly on the new Z10PE-D16 WS motherboard, with PCI Express slots designed to accommodate the power of multi graphics cards. Whether for professional graphics work, heavy duty multimedia or dedicated gaming, more than ample graphics power can be applied whenever needed.

Would be too convenient to have this motherboard with 8 PCIe slots as well (for 4 x GPUs)

Dual Xeon CPU Motherboard

USING ONLY ONE CPU ON DUAL CPU MOTHERBOARD?

Z10PE-D16 WS

The following table shows the supported CPUs for this motherboard

Intel Xeon E5-2620 v3 (2.4G,85W,L3:15M,6C,HT)

Single CPU configuration

You can refer to the following recommended memory population for a single CPU configuration.

Single CPU configuration (must be installed on CPU1)								
	DIMM							
	A2	A1	B2	B1	C2	C1	D2	D1
1 DIMM		✓						
2 DIMMs		✓		✓				
4 DIMMs		✓		✓		✓		✓
8 DIMMs	✓	✓	✓	✓	✓	✓	✓	✓

Dual CPU configuration

You can refer to the following recommended memory population for a dual CPU configuration.

Dual CPU configuration																
	DIMM (CPU1)								DIMM (CPU2)							
	A2	A1	B2	B1	C2	C1	D2	D1	E2	E1	F2	F1	G2	G1	H2	H1
2 DIMMs		✓								✓						
4 DIMMs		✓		✓						✓		✓				
8 DIMMs		✓		✓		✓		✓		✓		✓		✓		✓
12 DIMMs	✓	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓		✓
16 DIMMs	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2.5.4 PCI Express x16 slot (x16 link)

The on board PCIE 2 and 4 provide one x16 Gen3 link to CPU1, PCIE 5 and 6 provide one x16 Gen3 link to CPU2. These slots support VGA cards and various server class high performance add-on cards.

Only half of the PCIE slots and half of the memory slots work with one CPU if you do not populate both sockets.

See e.g.
servethehome.com

Limitations due to running dual socket mainboard with one CPU?

0

I installed a single Opteron CPU into a dual-socket mainboard Super Micro H8DGi-F and now I'm facing the problem that I can only use one PCI-e 16x slot, although there are 3 of them on the board. I've plugged in multiple video cards, but there's only one that is recognized by the OS.

In the manual (on page '1-9') I found a diagram showing "SLOT#6 PCIE (X16)" connected to "SOCKET #1", while the other two PCIE (X16) slots "#2" and "#4" are connected to "SOCKET #2"

superuser.com/questions

ars technica
openforum

Ars Technica > Forums > Hardware & Tweaking > CPU & Motherboard Technologia

dual-socket mobo with just one CPU

arstechnica.com

Motherboard XEON CPU XEON PHI

- The both suggested Asus motherboards should also be able to use with **XEON PHIS WHICH ARE A LOT PICKIER** with motherboards than NVIDIA GPUs.
  - If you want to keep the option to play with Xeon Phi, as the passive-cooled Xeon Phi (5110P, TDP 225W, you need the one with fan for desktop motherboards though) was selling for $200 at certain point.

Xeon Phi Knights Corner compatible workstation motherboards

POSTED BY VINCENT HINDRIKSEN ON 1 AUGUST 2015 WITH 3 COMMENTS

streamcomputing.eu

Motherboards with LGA 2011-3:

- Z10PA-D8: 2 x LGA 2011-3, 7x PCIe (x16x16)
- Z10PE-D16 WS: 2x LGA 2011-3, x PCIe (x16x16), 4 way x16
- Z10PE-D8 WS: x LGA 2011-3, x PCIe (x16x16), 4 way x16

STREAM COMPUTING
Performance Engineers

Will your motherboard work with Intel Xeon Phi?

Written on August 6, 2013 by Dr. Donald Kinghorn

https://www.pugetsystems.com/labs/hpc/Will-your-motherboard-work-with-Intel-Xeon-Phi-490/

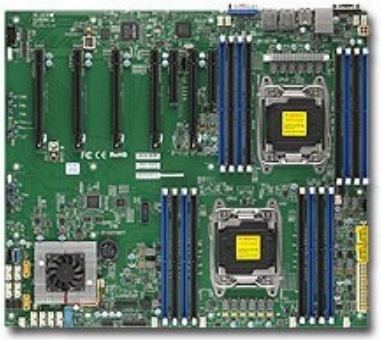
intel Developer Zone

CPU compatibility with Xenon Phi ?

Motherboard Dual XEON CPU SUPERMICRO "SERVER"

- If we are going with the dual-CPU path, the Z10PE-D16 WS allowed us to have only 3 GPUs

4 x GPUs
+ PCIe for networking



Key Features

- Supports 4-way Geforce SLI

1. Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 v4†/ v3 family; QPI up to 9.6GT/s
2. Intel® C612 chipset
3. Up to **2TB**† ECC 3DS LRDIMM , up to DDR4- **2400**†MHz ; 16x DIMM slots
4. 4x PCI-E 3.0 x16, 2x PCI-E 3.0 x8 (1 in x16), 1x PCI-E 2.0 x4 (in x8)
5. Intel® i350 Dual port GbE LAN
6. 10x SATA3 (6Gbps); RAID 0, 1, 5, 10
7. Integrated IPMI 2.0 and KVM with Dedicated LAN
8. 5x USB 3.0 ports, 4x USB 2.0 ports
9. HD Audio with optical S/PDIF

supermicro.com

Supermicro X10DRG-Q

by Supermicro
★★★★★ 2 customer reviews | 4 answered questions

Price: $502.00 + $16.50 shipping

~£390

Form Factor: Proprietary
Dimensions: 15.2" x 13.2" (38.6cm x 33.5cm)
Chipset: Intel® C612 chipset
CPU: Intel® Xeon® processor E5-2600 v4†/ v3 family (up to 160W TDP **)
Memory Capacity: 16x 288-pin DDR4 DIMM slots, up to 1 TB ECC RDIMM
Memory Type: 2400†/2133/1866/1600MHz ECC DDR4 SDRAM 72-bit
SATA: 10x SATA3 (6Gbps) ports **no M.2 SSD supported!**
PCI-Express: 4x PCI-E 3.0 x16 (double-width) slots
Supports 4-way Geforce SLI

Home › Reviews › IT/Datacenter › Motherboards

Supermicro X10DRG-Q (Intel C612) Workstation Motherboard Review

Supermicro's new X10DRG-Q workstation motherboard is built around HPC GPU applications using either GPU's, Quadra, Telsa, GRID, or Intel Xeon Phi. (NASDAQ:SMCI, NASDAQ:INTC)

By: William Harmon | **Motherboards** in **IT/Datacenter** | **Posted:** Nov 14, 2014 5:55 am

Comment | Email to a Friend | Font Size: A A
tweaktown.com

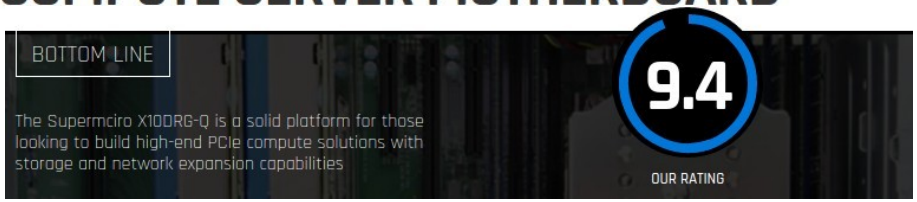
The Supermicro X10DRG-Q has the ability to run four PCIe expansion cards, which can be either GPU's, Quadro, Tesla, GRID, or Intel Xeon Phi co-processor cards. All of this is on one motherboard powered by dual Intel E5-2600 v3 processors, and up to 1TB of new DDR4 memory. The X10DRG-Q is a massive motherboard designed for HPC applications, or high-end workstation uses. Supermicro included all the bells and whistles on this motherboard, including storage options that can be run in RAID 0, 1, 5, and 10 configurations.

The focus of this motherboard is its use on the Supermicro GPU SuperWorkstation 7048GR-TR. We have the building blocks for a 7048GR-TR here in the lab, and we will be getting one up and running soon. We recommend getting the AOC-TBT-DSL5320 Thunderbolt Add-On Card for this motherboard because it will allow remote running of the machine, so it can be installed in a location separate from the user. These machines do make a fair amount of noise, so it is a good idea to install it in a separate room to keep the workplace less noisy.

Overall, this is a fantastic motherboard that has huge potential for HPC applications. One of the things we were not crazy about was the lack of iKVM when GPUs are installed, but the optional AOC-TBT-DSL5320 Thunderbolt Add-On Card takes care of this.ead more:

★ REVIEW

SUPERMICRO X10DRG-Q REVIEW – GPU COMPUTE SERVER MOTHERBOARD



servethehome.com, by PATRICK KENNEDY, JANUARY 21, 2015

Conclusion

We have been wanting to take one of Supermicro's GPU compute platforms out for a spin and the Supermicro X10DRG-Q certainly exceeded our expectations.

The platform booted up the first time in every configuration we tried. Mellanox FDR Infiniband cards, 40GbE Intel Fortville cards, RAID controllers, Xeon Phi's, NVIDIA Quadro cards we tested all booted without issue. For those seeking a quad Xeon Phi or quad Tesla/ Quadro solution, the Supermicro X10DRG-Q should be atop lists.

Supermicro Chassis NON-STANDARD FORM FACTOR

Key Features

1. 1620W Redundant Power Supplies **Platinum Level (94%)**
2. 4x dedicated Power Connectors (6pin + 8pin) for High-end GPUs
3. 8x 3.5" Hot-swap SAS/SATA HDDs
4. 11x Full-height, Full-length expansion slots optimized for 4x double width GPU solution
5. Advanced Fan Speed Control
6. Power Switch, Reset Switch & 5 LED Indicators; 2x Front USB 2.0 Ports
7. Highest Quality Chassis Built - Superb Components, Fans, & Power Supply
8. Full SES2 support is available with SAS motherboards and other compatible components.

Note: Image above may show a varied configuration or optional parts, Please refer to parts list for standard parts included.

SuperChassis 747TQ-R1620B

Products Chassis 4U [SC747TQ-R1620B]

Click to open expanded view

Supermicro CSE-747TQ-R1620B 1620W 4U Tower/Rackmount Server Chassis

by Supermicro

Be the first to review this item

Price: $1,069.88 & **FREE Shipping**

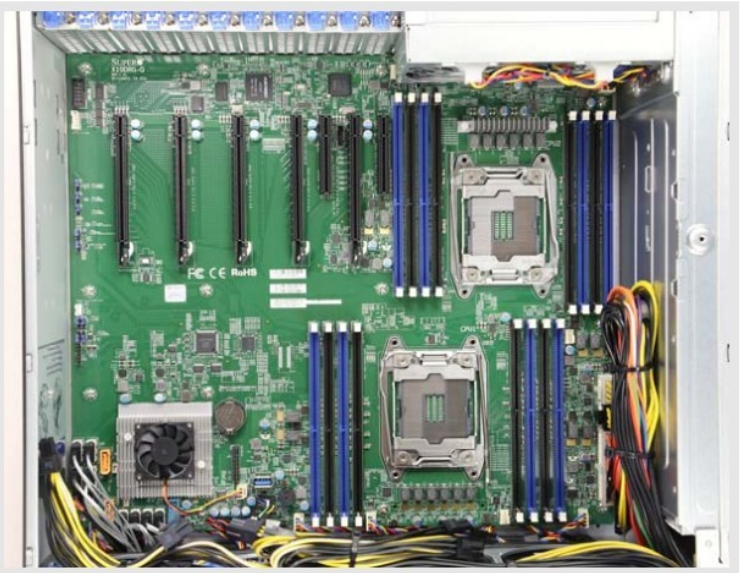
Item is eligible: **No interest if paid in full within 12 months** with the Amazon.com Store Card. Apply now

Only 1 left in stock.

This item ships to **LONDON, United Kingdom**.

Ships from and sold by CE Showroom.

- Form Factor: Tower/4U chassis support for DP and UP ATX, E-ATX motherboards: up to size 15.2-Inch x 13.2-Inch
- Expansion Slots: Capable of housing 11x Full-Height, Full-Length expansion cards
- SAS/SATA Backplane: SAS/SATA Hard Drive Backplane with SES2
- System Monitoring: Chassis intrusion switch
- Power Supply: 1620W high-efficiency redundant power supply with PMBus



Supermicro X10DRG-Q overview

We are using a Supermicro **SC747TQ-R1620B**chassis to house the X10DRG-Q due to its size, and highly recommend that purchasers of the motherboard do the same. servethehome.com

Power Supply	
1620W high-efficiency redundant power supply w/ PMBus	
AC Input	▪ 1000W Output @ 100-120V, 12-10A, 50-60Hz ▪ 1200W Output @ 120-140V, 12-10A, 50-60Hz ▪ 1620W Output @ 180-240V, 10.5-8A, 50-60Hz
DC Output	▪ 1000W: +12V/84A; +5Vsb/4A ▪ 1200W: +12V/100A; +5Vsb/4A ▪ 1620W: +12V/150A; +5Vsb/4A
Certification	80 PLUS PLATINUM 94% **Platinum Certified** [Test Report]

IBM “PowerAI Server”

IBM and Nvidia team up to create deep learning hardware

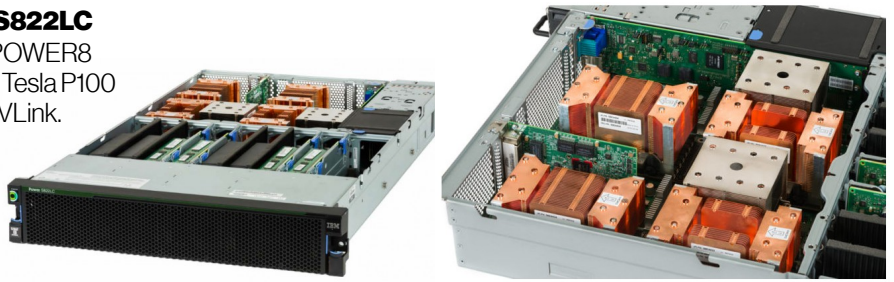
http://venturebeat.com/2016/11/14/ibm-and-nvidia-team-up-to-create-deep-learning-hardware/

DEAN TAKAHASHI NOVEMBER 14, 2016 2:00
AMTAGS: DEEP LEARNING NEURAL NETWORKS, IBM, IBM POWERAI, NVIDIA, POWER, TOP-STORIES

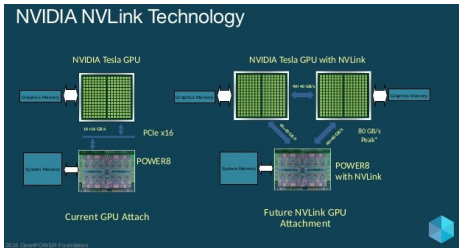
A new software toolkit available today called IBM PowerAI is designed to run on the recently announced IBM server built for artificial intelligence that features Nvidia NVLink technology optimized for IBM’s Power Architecture.

The hardware-software solution provides more than **two times the performance** of comparable servers with four graphics processing units (GPUs) running AlexNet with Caffe. The same four-GPU Power-based configuration running AlexNet with BVLC Caffe can also **outperform 8-M40 GPU-based x86 configurations**, making it the **world’s fastest commercially available enterprise systems platform** on two versions of a key deep learning framework, the companies said.

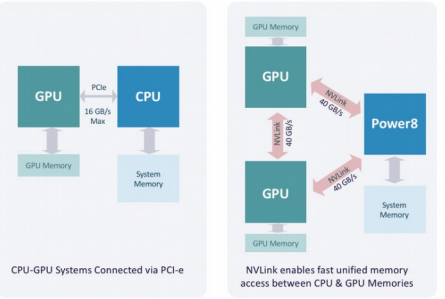
IBM Power System S822LC incorporates two IBM POWER8 CPUs with four NVIDIA Tesla P100 GPUs connected via NVLink.



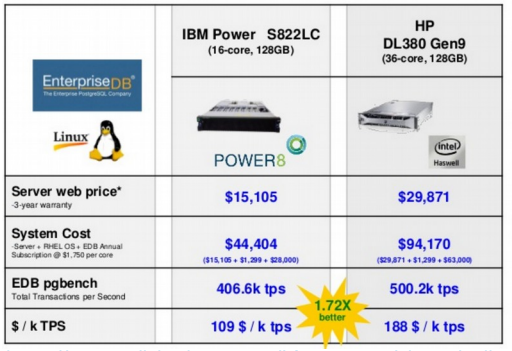
https://blogs.nvidia.com/blog/2016/09/08/ibm-servers-nvlink/
https://www.microway.com/product/openpower-gpu-server-nvidia-tesla-p100-nvlink-gpus/

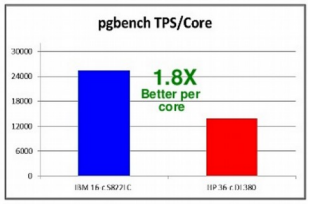


http://www.slideshare.net/mandieq/openpower-solutions-overview-session-from-ibm-techu-rome-april-2016



http://www.kinetica.com/partner/ibm/





1.72X Better Price-performance

http://www.slideshare.net/KangarootLinux/9-ibm-power-open16

Asus Server Motherboard 8 X GPUS

- If desktop workstation is not enough for you, go for the rack-mountable server, and build that yourself as well.

ASUS 1st 3U 8GPU Super-Computer with Extreme Density and Hybrid Computing Power

After the long hard working in HPC field, ASUS now presents the GPU server ESC8000 G3 supporting up to **eight double deck GPU cards** as well as optimized thermal design for both CPUs and GPUs. ESC8000 G3 is the latest supercomputer based on Intel® Xeon® processor E5-2600 v3 product family, featuring:

- Front parallel redundant fan placement and dedicated air-tunnel for individual GPUs,
- 6 hot-swappable 2.5" SATA HDD/SSD bays,
- 2+1 80 PLUS Platinum 1600W CRPS ,
- 8 PCI-E Gen3 x16
- 2 PCI-E Gen3 x8 expansion slots.

ESC8000 G3 is targeted to hit 20 Tera floating points with the latest generation GPU cards and achieve an outstanding performance result in Top500 and Green500.

ESC8000 G3 is equipped with dual Intel® Xeon® processor E5-2600 v3 product family CPUs, 24 DDR4 DIMMs and support up to eight dual-slot GPGPU; thus delivers high density computing power with scalable expansion capability and an intelligent thermal solution, making it an ideal choice for the applications in the field of life and medical science, engineering science, financial modeling and virtualization.

Asus ESC8000_G3

For deep learning use, now we can use all the 8 PCI-E slots for **8 GPU cards** doubling performance compared to desktop motherboards.

With desktop motherboards, we don't have **any PCI-E slots left** if we go for 4 GPUs which you may want to use for faster **network cards** (e.g. for NVIDIA GPUDirect RDMA, 40Gb/s).

ESC8000 G3

3U Rackmount Server User Guide

pugetsystems.com

Asus Esc8000 G3 Barebone System - 3u Rack-Mountable - Intel C612 Chipset - Socket R3 (Lga2011-3) - 2 X Processor Support - 1.50 Tb Ddr4 Sdram Ddr4-2133/Pc4-17000 Maximum Ram Support - Serial Ata/600 Raid Supported Controller - Aspeed Ast2400 32 Mb In

Part Number:	ESC8000 G3
Usually Ships:	Not in stock
Stock:	No | Email Alert !
	0 Click for Availability

Your Price: **$3,200.00**

Unboxing NVIDIA Geforce GTX Titan X + Asus Server ESC8000 G3 !!!

youtube.com

Asus ESC8000 G3 SPECIFICATIONS

ESC8000 G3

The following table shows the supported CPUs for this motherboard Click here to search other motherboards.

CPU	Validated since PCB	Validated since BIOS	Note
E5-1603 V4(2.80GHz,140W,L3:10M)		3202 GO	
E5-1607 V4(3.10GHz,140W,L3:10M)		3202 GO	
E5-1620 V4(3.50GHz,140W,L3:10M,4C,HT)		3202 GO	
E5-1630 V4(3.70GHz,140W,L3:10M,4C,HT)		3202 GO	
E5-1650 V4(3.60GHz,140W,L3:15M,6C,HT)		3202 GO	
E5-1660 V4(3.20GHz,140W,L3:20M,8C,HT)		3202 GO	
E5-2603 V4(1.70GHz,85W,L3:15MB,6C)		3202 GO	
E5-2609 V4(1.70GHz,85W,L3:20MB,8C)		3202 GO	
E5-2620 V4(2.10GHz,85W,L3:20M,8C,HT)		3202 GO	
E5-2623 V4(2.60GHz,85W,L3:10M,4C,HT)		3202 GO	
E5-2630 V4(2.20GHz,85W,L3:25M,10C,HT)		3202 GO	
E5-2630L V4(1.80GHz,55W,L3:25M,10C,HT)		3202 GO	
E5-2637 V4(3.50GHz,135W,L3:15M,4C,HT)		3202 GO	
E5-2640 V4(2.40GHz,90W,L3:25M,10C,HT)		3202 GO	
E5-2643 V4(3.40GHz,135W,L3:20M,6C,HT)		3202 GO	

asus.com

Processor / System Bus	2 x Socket R3 (LGA 2011-3) Intel® Xeon® processor E5-2600 v4 product family (145W) Intel® Xeon® processor E5-2600 v3 product family(145W) QPI 6.4 / 8.0 / 9.6 GT/s *Refer to www.asus.com for CPU Support list
Core Logic	Intel® C612 PCH
Memory	**Total Slots :** 24 (4-channel per CPU, 12 DIMM per CPU) **Capacity :** Maximum up to 1536GB LRDIMM **Memory Type :** DDR4 2400 /2133 /1866/1600/1333 RDIMM*1 DDR4 2400 /2133 /1866/1600/1333 LRDIMM*1 DDR4 2400 /2133 /1866/1600/1333 NVDIMM*1 **Memory Size :** 32GB, 16GB, 8GB, 4GB RDIMM 64GB, 32GB LRDIMM * Please refer to ASUS server AVL for the latest update
Expansion Slots	Total: 10 Full-length/Full-height 8 x PCI-E 3.0 x16 (x16 Link) + 1 x PCI-E 3.0 x16 (x8 Link) Half-lenth/Low-profile 1 x PCI-E 3.0 x8 (x8 Link)
Dimensions	759mm x 447mm x 130.6mm (3U)
Form Factor	3U
Weight	29.2kg (exluding CPU/memory/HDD/add-on cards)
Power Supply	2+1 Redundant 1600W 80PLUS Platinum Power Supply; 1600W: 100-127/200-240 Vac, 25.8/19 A, 47-63 Hz, Class I

Storage	**SATA Controller :** Intel® C612 6 x SATA3 6Gb/s ports 1 x M.2 connector (2280/2260/2242) Intel® Rapid Storage Technology Enterprise(RSTe) (For Windows Only) (Support Software RAID 0, 1, 5, 10) LSI® MegaRAID (For Linux/Windows) (Support Software RAID 0, 1, 10) **SAS Controller :** ASUS PIKE II 3008 8-port SAS HBA card*2 ASUS PIKE II 3108 8-port SAS HW RAID card*2
HDD Bays	6 x Hot-swap 2.5" HDD Bays
Networking	2 x Intel® I210AT + 1 x Mgmt LAN
InfiniBand	Optional kits: PEM-FDR PEB-10G/57840-2S PEB-10G/57811-1S
Graphic	Aspeed AST2400 with 32MB VRAM
Front I/O Ports	1 x COM port 2 x USB 3.0 ports 2 x USB 2.0 ports 1 x VGA port
Rear I/O Ports	3 x RJ-45 ports (One for ASMB8-iKVM)
OS Support	Windows® Server 2012 R2 Windows® Server 2012 Windows® Server 2008 Enterprise R2 SP1 64-bit RedHat® Enterprise Linux SuSE® Linux Enterprise Server CentOS Ubuntu VMware Citrix XenServer *(Subject to change without any notice)
Management Solution	ASWM Enterprise Default 1 x ASMB8-iKVM for KVM-over-Internet

https://www.asus.com/uk/Commercial-Servers-Workstations/ESC8000_G3/specifications/

Budget change WITH SERVER MOTHERBOARDS

- **"CSE-747TQ-R1620B chassis + X10DRG-Q MB"** comes with power source so we can reduce the cost of workstation MB (~£500) , case (£100), and power supply (~£250)
  - 8x 3.5" Hot-swap SAS/SATA HDDs
  - 11x Full-height, Full-length expansion slots optimized for 4x double width GPU solution

Supermicro CSE-747TQ-R1620B 1620W 4U Tower/Rackmount Server Chassis
by Supermicro
Be the first to review this item

Price: $1,069.88 & **FREE Shipping**

Item is eligible: **No interest if paid in full within 12 months** with the Amazon.com Store Card. Apply now

Only 1 left in stock.

Supermicro X10DRG-Q

by Supermicro
2 customer reviews
4 answered questions

Price: $502.00 + $16.50 shipping

~£390

~(£1,150 - 850), thus **£300 more expensive** build than the "consumerdual XEON options"

~ £7,800 + £300
= ~£8.100

~£8,100 + £920 =
~£9,000 4th Titan-X

- **ESC8000 G3** comes with a ASUS Z10PG-D24 Server Board, and the power source so we can reduce the cost of workstation MB (~£500), case (£100) and power supply (~£250)
  - Excluding the 3.5" HDDs which have no space in the case, and those need an external case or a 3U/4U NAS server. The case/MB itself can house 6 x 2.5 Sata III SSDs, and one M.2 SSD.

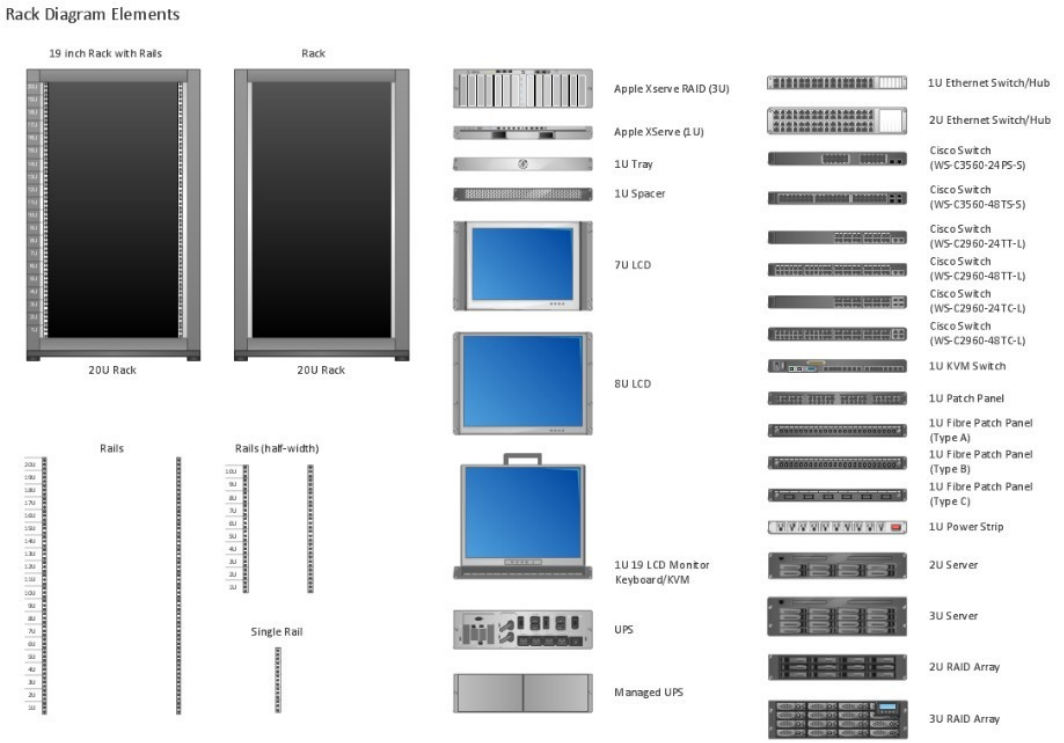
Shop on Google Sponsored

Asus ESC8000 G3 (ASMB8-IKVM) 3U Rack Server
£3,328.96 - Ebuyer.com
Free shipping

~(£3,330 - 850), thus **£2.450 more expensive** build than the "consumer options"

Server Rackmount Chassis RACK UNITS

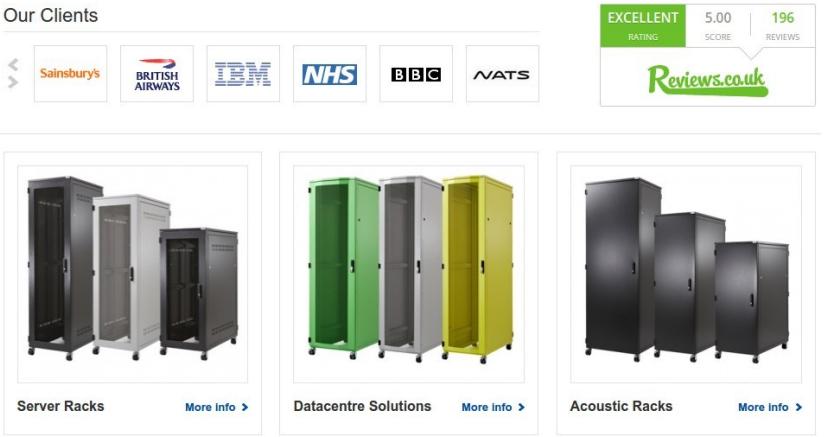
Wikipedia:"A **rack unit** (abbreviated **U** or RU) is a unit of measure defined as 1.75 inches (44.45 mm). It is most frequently used as a measurement of the overall height of 19-inch and 23-inch rack frames, as well as the height of equipment that mounts in these frames, whereby the height of the frame or equipment is expressed as multiples of rack units. For example, a typical full-size rack cage is 42U high, while equipment is typically 1U, 2U, 3U, or 4U high.



http://www.conceptdraw.com/examples/rack-diagram

6u 450mm Deep Wall Mounted Data Cabinet. £74.40 Inc VAT.

cannontech.co.uk



http://www.rackcabinets.co.uk/, e.g.

SR9-6-12	9U	600mm x 690mm x1200mm £594.00 (inc VAT)
SR30-8-10	30U	800mm x 1625mm x 1000mm £576.00 (inc VAT)

RAM DDR4 LATENCY AND CLOCK RATE

Figure 1

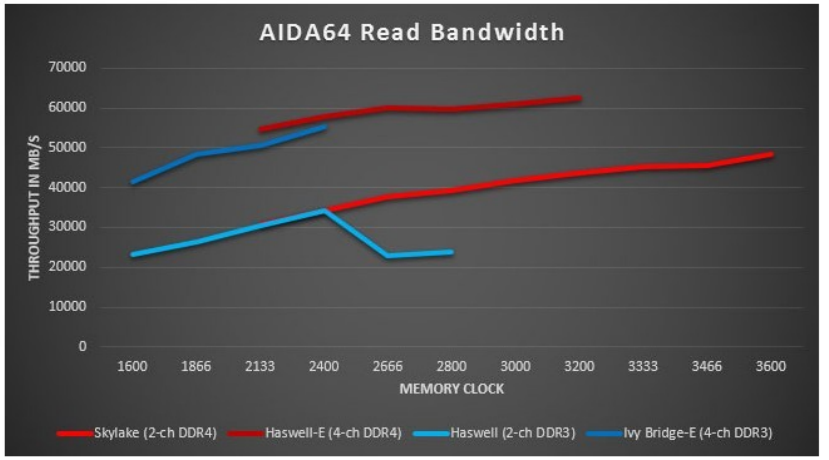
SPEED VS. LATENCY AS MEMORY TECHNOLOGY HAS MATURED (INDUSTRY STANDARDS)

TECHNOLOGY	MODULE SPEED (MT/s)	CLOCK CYCLE TIME (ns)	CAS LATENCY (CL)	TRUE LATENCY (ns)
SDR	10E	8.00	3	24.00
SDR	133	7.50	3	22.50
DDR	335	6.00	2.5	15.00
DDR	40B	5.00	3	15.00
DDR2	667	3.00	5	15.00
DDR2	800	2.50	6	15.00
DDR3	1333	1.50	9	13.50
DDR3	160B	1.25	11	13.75
DDR4	1866	1.07	13	13.93
DDR4	2133	0.94	15	14.06
DDR4	2400	0.83	17	14.17
DDR4	2666	0.75	18	13.50

In the history of memory technology, as speeds have increased, clock cycle times have actually decreased, resulting in lower true latencies as technology has matured, even though there are more clock cycles to complete.

http://www.anandtech.com

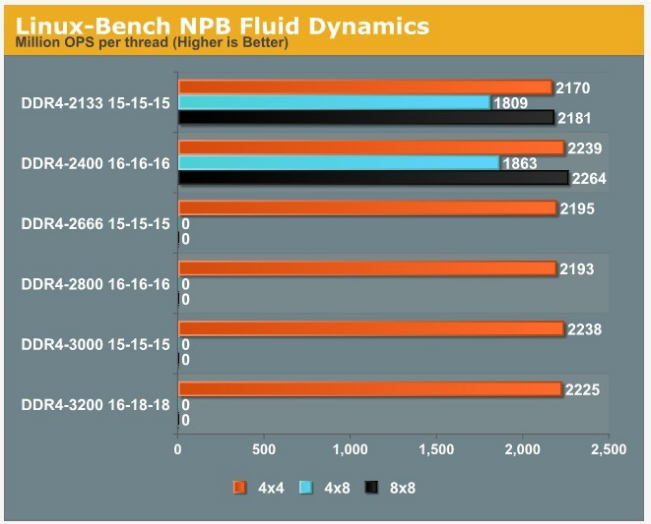
Read Speed

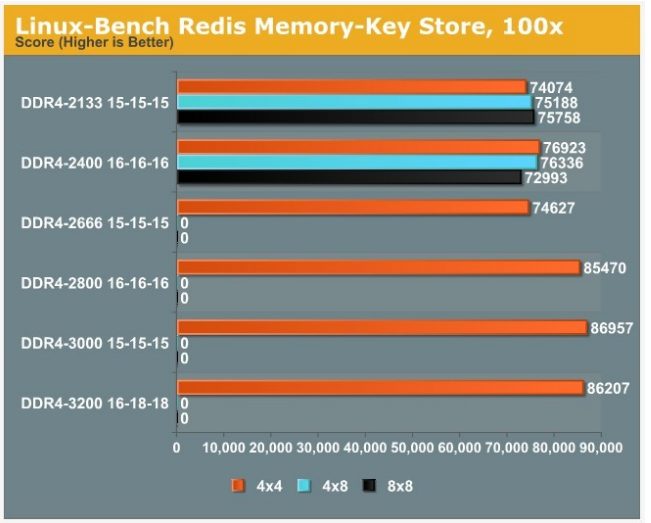


corsair.com

2,133 MHz (PC4-17000) seems like a good choice in practice, but the DDR4 EEC 2,400 MHz (PC4-19200) is very close price-wise to 2,133 MHz (£20 more per 64 GB) so we might as well buy the faster RAM

Crucial - DDR4 - 64 GB : 4 x 16 GB - DIMM 288-pin - 2400 MHz / PC4-19200 - CL17 - 1.2 V - registered – ECC – **Price £320**





256 GB of RAM in Zlateski et al. (2016)

> The benchmarks are performed on two machines. The first machine is a 4-way Intel Xeon E7 8890v3 with total of 72 cores (144 hyper-threads), 256GB of RAM and a Titan X GPU (with 12GB on-board RAM). The second machine is an Amazon EC2 instance with 32 virtual cores and 244GB of RAM (r3.8xlarge). The second machine is included as it is more readily available.

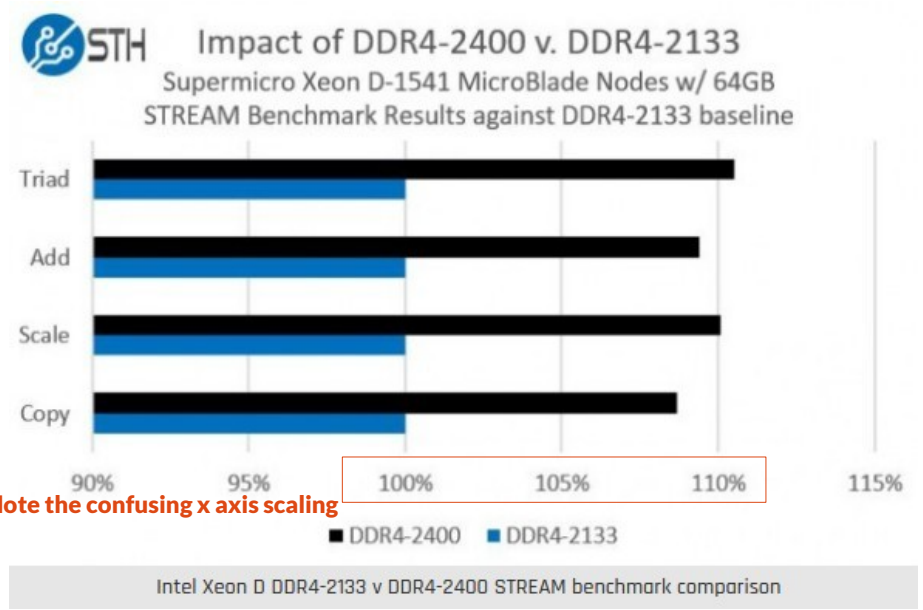
DDR4 Haswell-E Scaling Review: 2133 to 3200 MHz

http://www.anandtech.com/show/8959/ddr4-haswell-e-scaling-review-2133-to-3200-with-gskill-corsair-adata-and-crucial/5

RAM DDR4 PERFORMANCE CONTINUED

Benchmarking DDR4-2133 v. DDR4-2400 using STREA and XEON D (low-power system on a chip)

"For this test we are using the industry standard STREAM benchmark. STREAM is a benchmark that needs virtually no introduction. It is considered by many to be the de facto memory performance benchmark. Authored by John D. McCalpin, Ph.D. it can be found at http://www.cs.virginia.edu/stream/ and is very easy to use."



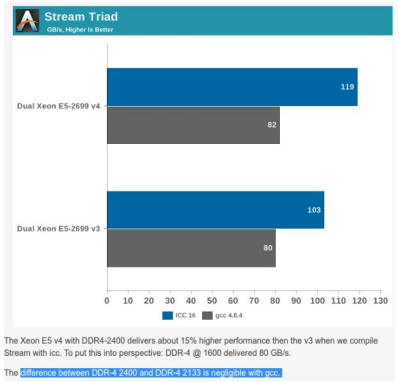
Note the confusing x axis scaling

Intel Xeon D DDR4-2133 v DDR4-2400 STREAM benchmark comparison

http://www.servethehome.com/ddr4-2133-versus-ddr4-2400-stream-benchmarks-using-xeon-d/

From a raw MHz gain perspective, we would **expect to see a 12.5% increase** in performance. This gain is tempered by the increase in latency from 15-15-15 (DDR4-2133) to 17-17-17 (DDR4-2400) on the higher speed modules. Every time we see a DDR frequencey bump, we see the latency figures creep up as well. We ran the STREAM benchmark on each node over a 12 hour period to ensure we had a repeatable result set and we threw out any results that were more than 3 sigma off of the mean (assuming these were benchmark tool anomolies.) As you can see from the above, one can assume approximately a **9.5-10.5% improvement in memory** bandwidth for the Xeon D platform, and we would expect from other similar platforms.

"Memory-optimized" applications will see the bandwidth increase



anandtech.com

White Paper
FUJITSU Server PRIMERGY
Memory Performance of Xeon E5-2600 v4 (Broadwell-EP) based Systems

sp.ts.fujitsu.com

RAM DDR4 ECC

Hacker News new | comments | show | ask | jobs | submit

▲ Why use ECC? (danluu.com)

188 points by benkuhn 261 days ago | hide | past | web | 95 comments | favorite

https://news.ycombinator.com/item?id=10638324

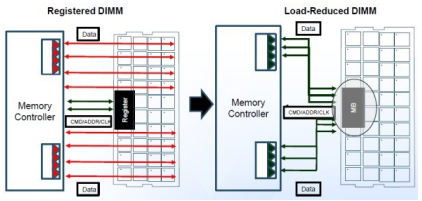
nextos 261 days ago [-]

This is a big dilemma I have. I'm trying to build a workstation similar to Nvidia's reference design for deep learninghttps://developer.nvidia.com/devbox I will be doing deep learning and other ML GPU-powered tasks. Plus some long running high-memory I/O intensive tasks. Note that Nvidia's build does not employ ECC RAM. And it's a quite expensive machine. Mine will be only a fraction of the cost ($4.5k), with just one Titan X. It's possible to afford a Xeon, but this comes at the compromise of buying slower hardware. What shall I do? Intel's segmentation of the market, limiting the amount of RAM you can use in regular CPUs and removing support for ECC sucks.

Alupis 261 days ago [-]

They likely don't use ECC ram because it's a Dev box targeted at development and testing... ie. it's not going to need to be up and running for a large stretch of time and/or data corruption is acceptable. Otherwise, their reference system is no different from a "high end" gaming rig (ie. not a "server")

RDIMMs add a register, which buffers the address and command signals.The integrated memory controller in the CPU sees the register instead of addressing the memory chips directly. As a result, the number of ranks per channel is typically higher: the current Xeon E5 systems support up to eight ranks of RDIMMs. That is four dual ranked DIMMs per channel (but you only have three DIMM slots per channel) or two Quad Ranked DIMMs per channel. If you combine quad ranks with the largest memory chips, you get the largest DIMM capacities. For example, a quad rank DIMM with 4Gb chips is a 32GB DIMM (4 Gbit x 8 x 4 ranks). So in that case we can get up to 256GB: 4 channels x 2 DPC x 32GB. Not all servers support quad ranks though.



LRDIMMs can do even better. Load Reduced DIMMs replace the register with an Isolation Memory Buffer (iMB™ by Inphi) component. The iMB buffers the Command, Address, and data signals.The iMB isolates all electrical loading (including the data signals) of the memory chips on the (LR)DIMM from the host memory controller. Again, the host controllers sees only the iMB and not the individual memory chips. As a result you can fill all DIMM slots with quad ranked DIMMs. In reality this means that you get 50% to 100% more memory capacity.

http://www.anandtech.com/show/6068/lrdimms-rdimms-supermicros-latest-twin/2

lasagne-users › **How important is ECC (in GPU RAM) for Neural Networks?**

I've just found two interesting articles, which are slightly related:

+ http://blog.codinghorror.com/to-ecc-or-not-to-ecc/
+ https://storagemojo.com/2012/10/23/dram-errors-soft-and-hard/

Key points I take out of it and your answers are:

+ Sometimes ECC is cheap to get. Then one should take it.
+ ECC is crucial when one runs big calculations long enough so (1) errors become more likely (2) it becomes VERY annoying / expensive to run the calculation again
+ Neural networks can deal with noisy data, so bitflips should not be a problem
+ Having Maxwell architecture is much more important than having ECC

CODING HORROR
programming and human factors

19 Nov 2015

To ECC or Not To ECC

https://blog.codinghorror.com/to-ecc-or-not-to-ecc/

Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors

Yoongu Kim[1] Ross Daly* Jeremie Kim[1] Chris Fallin* Ji Hye Lee[1]
Donghyuk Lee[1] Chris Wilkerson[2] Konrad Lai Onur Mutlu[1]
[1]Carnegie Mellon University [2]Intel Labs

doi:10.1109/ISCA.2014.6853210
github.com/CMU-SAFARI/rowhammer

UNBUFFERED VERSUS REGISTERED ECC MEMORY – DIFFERENCE BETWEEN ECC UDIMMS AND RDIMMS

39

Downsides of ECC RAM

ECC is designed to be more stable than traditional RAM, and our failure records show that this is indeed the case. However, there are a few downsides to using ECC RAM. The first, and most obvious, is that not every computer can use ECC memory. Most server and workstation motheboards require ECC RAM, but the majority of desktop systems either won't work at all with ECC RAM or the ECC functionality will be disabled.

Second, due to the additional memory chip and the inherently more complex nature of ECC RAM, it costs more than non-ECC RAM. The amount varies, but you should expect to pay roughly 10-20% more depending on the size of the memory stick. The larger the stick, the higher the price premium.

Finally, ECC RAM is slightly slower than non-ECC RAM. Many memory manufacturers say that ECC RAM will be roughly 2% slower than standard RAM due to the additional time it takes for the system to check for any memory errors. To verify this, we examined multiple benchmarks that we run on each system we produce. By using comparable CPUs (For example: Intel Core i7 477† 3.5GHz Quad Core 8MB versus Intel Xeon E3-1275 V3 3.5GHZ Quad Core 8MB) we found that this 2% estimate to be roughly correct. Our own benchmarks showed a performance hit ranging from .72 to 2.2% which, given normal testing deviations, is right in line with the 2% estimate.

pugetsystems.com

by PATRICK KENNEDY
MARCH 9, 2011

Data integrity

Bernd Panzer-Steindel, CERN/IT
Draft 1.3 8. April 2007

We have 44 reported memory errors (41 ECC and 3 double bit) on ~1300 nodes during a period of about 3 month. The memory vendors quote a Bit Error Rate of 10-12 for their memory modules. .. Thus the observed error rate is **4 orders of magnitude lower than expected** for the single bit errors, while one would expect no double bit errors.

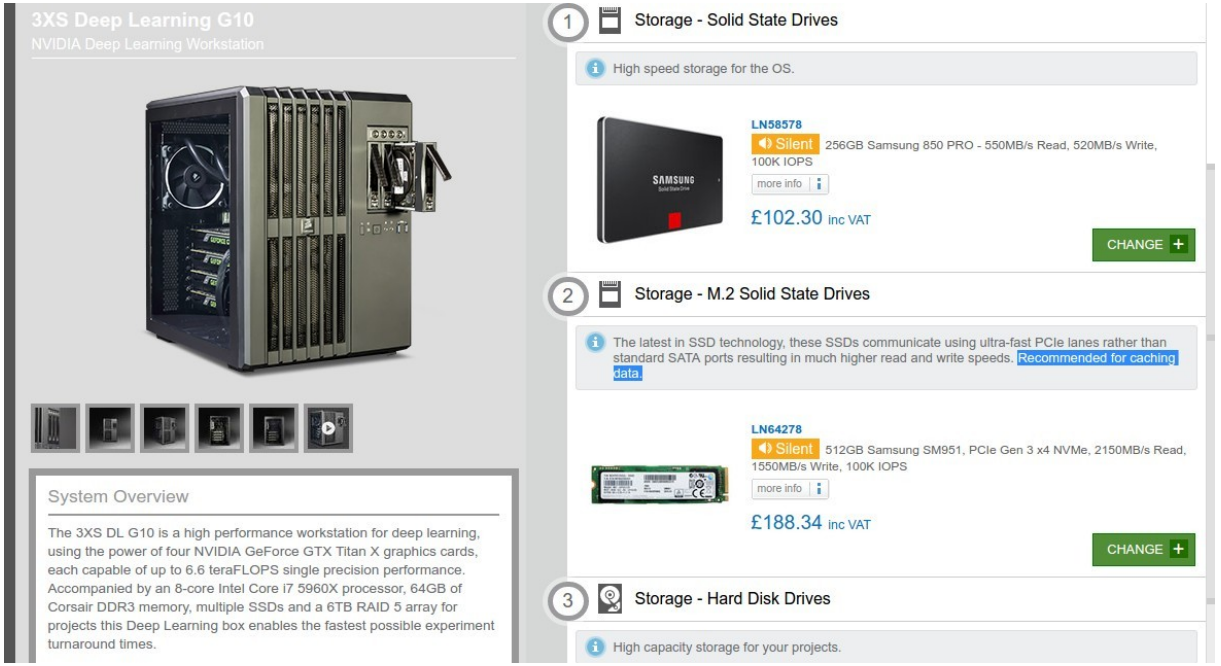
Hard Drives

Three types of HDDs:

1) Sata 6Gb/s SSD
   - For OS (e.g. Ubuntu 14.04)

2) M.2 SSD
   - The fastest (seek times, and write/read speeds). Use for caching data sets read from "spinning HDDs" → improves RAID performance

3) "Spinning old school" HDDs
   - For data storage



https://www.scan.co.uk/3xs/configurator/nvidia-deep-learning-box--3xs-g10

3

Western Digital Red Pro

Prices (Aug 2016), 7200 rpm

2 TB	£120	£60/TB
3 TB	£135	£45/TB
4 TB	£199	£50/TB
6 TB	**£213**	**£36/TB**
8 TB	£430	£54/TB

1

Samsung 850 Pro SSD

Prices (Aug 2016), Sata III

1024 GB = 1 TB

120GB	£70	£597/TB
250GB	£110	£451/TB
500GB	£175	£358/TB
1 TB	**£337**	**£337/TB**
2 TB	£734	£367/TB

HDD Schematics

ALONG WITH AVERAGE READ SPEEDS

SATAIII
SPINNING STORAGE

~180 MB/S

SATAIII SSD

~550 MB/S

SATA III
MAX 600 MB/s

M.2 SSD

~2.2 GB/S

~55 GB/S
DDR4 2133
PC4-17000

IF YOU CAN FIT WHOLE OF YOUR DATA TO DDR-4 RAM, YOU WILL REDUCE THE **DISK LATENCIES** A LOT

In practice your datasets might be so large, that even the proposed 256 GB of RAM might run out of fast.
Then you can take advantage of **prefetching** queue / scheduling, as found from TensorFlow (Threading and Queues). for example.

Tim Dettmers: "The memory bandwidth of your RAM determines how fast a mini-batch can be overwritten and allocated for initiating a GPU transfer, but the next step, **CPU-RAM-to-GPU-RAM is the true bottleneck** – this step makes use of direct memory access (DMA). As quoted above, the memory bandwidth for my RAM modules are 51.2GB/s, but the DMA bandwidth is only 12GB/s!

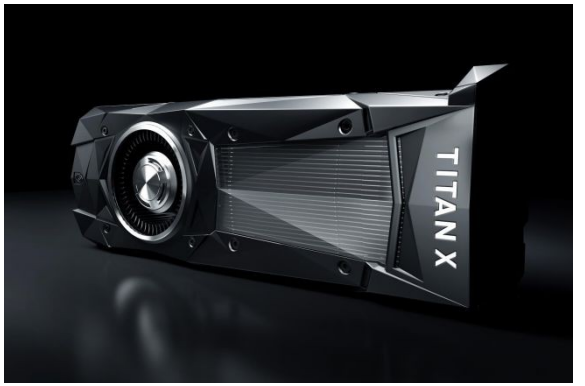
If we take the batch size to be 128 and the dimensions of the data 244x244x3 that is a total of roughly 0.085 GB ($4*128*244^2*3*1024^{-3}$). With an ultra-slow **memory** we have 6.4 GB/s, or in other terms 75 mini-batches per second! So with asynchronous mini-batch allocation **even the slowest RAM will be more than sufficient** for deep learning.

HDD We have in the case of the Alex's ImageNet convolutional net 0.085GB every 0.3 seconds, or 290MB/s if we save the data as 32 bit floating data. If we however save it as jpeg data, we can compress it 5-15 fold bringing down the required read bandwidth to about 30MB/s.

~60 GB/S

TITAN X MAXWELL
GDDR5X
Clock: 10 Gbps

PCI-E V.3
~15.75 GB/S



250 GB per OS (SATA III SSD or the M.2 SSD if you want extra responsivity to your OS use) should be comfortable (500 GB if you for example dual-boot with Windows)

LOCAL RAID

CASES

- Fractal Define XL R2, **8 x** 3.5” HDD bays
- Supermicro SC747TQ-R1620B, **8 x** 3.5” HDD bays, SC747TQ-R1620B MB supports 10 x SATA III drives

6 TB
5,400rpm

WD 6 TB NAS Desktop Hard Disk Drive (Intellipower SATA 6 Gb/s 64 MB Cache) - 3.5 inch, Red
by Western Digital
1,173 customer reviews | 92 answered questions

RRP: £213.26
Price: £195.71 & **FREE Delivery** in the UK. Details
You Save: £17.55 (8%)

In stock.

Want it Saturday, 18 June? Order it within **8 hrs 57 mins** and choose **Priority Delivery** at checkout. Details

Sold by shopherellc and Fulfilled by Amazon. Gift-wrap available.

18 TB (raw) 24 TB (raw)
£3x200 - £3x430
5,400rpm → 7,200rpm

12 TB or 18 TB of usable space (RAID 5)

- WD Red is designed specifically for home and small office NAS systems and PCs with RAID
- NAS compatible
- NASware 3.0 enables seamless integration
- Lower operating temperatures
- More reliable solution for all NAS environments

WD 8TB Red Pro 3.5" SATA 3 NAS HDD/Hard Disk Drive
WD8001FFWX

8TB WD Red Pro WD8001FFWX, 3.5" NAS 24x7 HDD, SATA III - 6Gb/s, 7200rpm, 128MB Cache, NCQ, OEM

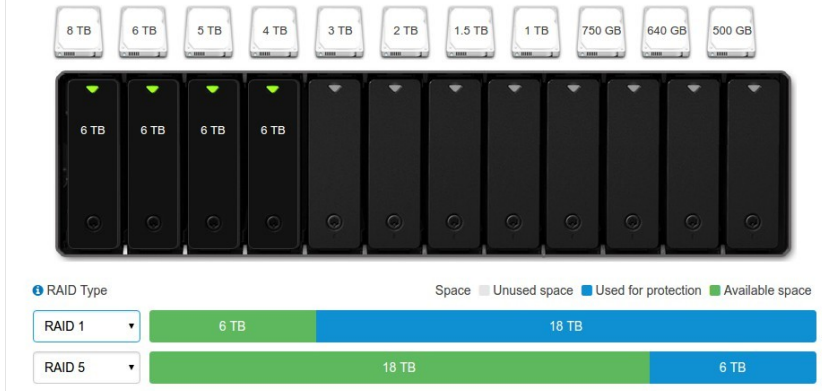
8 TB

Scan Code: LN73117
Manufacturer Code: WD8001FFWX
Scan Rating: High End
Write the first review
Follow this product
Be the first to ask a question.

7,200rpm

In Stock
Delivered to you on Tue 9th Aug
£358.32 Ex VAT
£429.98 Inc VAT
ADD TO BASKET
Need Help? CALL ME | LIVE CHAT
Delivery
Delivered by DPD £5.48
Tue 09 Aug. Delivered to your specified address. Receive SMS with one-hour delivery window.
Collect Instore Free
Place your order online and collect from our Bolton store with Q-Collect.
Weekend, timed and European delivery options are available at

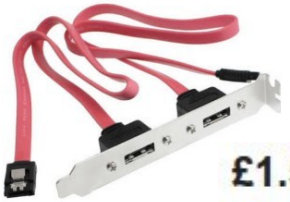
RAID Calculator
https://www.synology.com/en-global/support/RAID_calculator



Higher £/GB but if you really want a lot of storage, and have limited space

Hot-Swap Enclosure

- If you don't enough HDD spaces in your case (or you use Asus ESC8000 rack-mounted server with no 3.5” HDD bays)



Dual Port SATA Serial ATA Cable to eSATA Bracket Adapter

ebay.co.uk

£1.55

Free shipping

The eSATA interface delivers fast SATA III speeds up to 6Gbps, and allows you to access four drives via a single eSATA connection (eSATA port multiplier required). It is compatible with most of RAID controller card that has eSATA port-multiplier interface, which gives you an option to setup a hardware RAID externally, perfect for system that has limited space. This tiny enclosure can take up to 32TB total capacity with largest capacity drives available on the market (8 TB), which is more than enough space for most of the applications.

http://www.icydock.com/goods.php?id=172

RAID config in Ubuntu

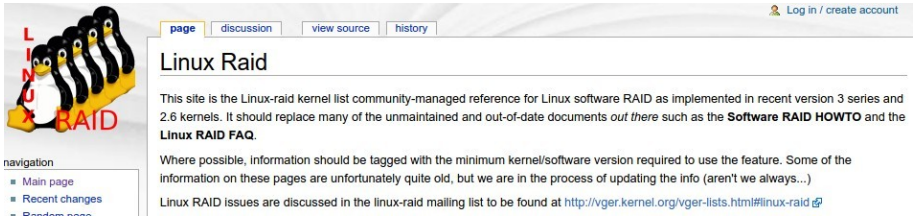
- See Dmytro Prylipko's article on building a “DIY: Deep Learning DevBox” with useful tips for RAID5 configuration in Ubuntu 14.04.

DIY: Deep Learning DevBox

Published on March 14, 2016
https://www.linkedin.com/pulse/diy-deep-learning-devbox-dmytro-prylipko

Dmytro Prylipko | Follow
Machine Learning Engineer at BuddyGuard GmbH

13 1 0



https://raid.wiki.kernel.org/index.php/Linux_Raid

Overview

There is an Overview section that is based on the RAID HowTo, covering the following:

Why RAID?
Devices
Hardware issues
RAID setup
Detecting, querying and testing
Tweaking, tuning and troubleshooting
Reconstruction
Recovering a failed software RAID
Growing
Performance
Related tools
Partitioning RAID / LVM on RAID

The document is sprinkled with references to the deprecated (since 2003) raidtools which are being gradually removed. Anything mentioning mkraid, raidtab or raidtools should be fixed.

“Consumer” Cases | BIG, QUIET, STAY “COOL”

Define XL R2 Black Pearl

Specifications

- ATX, Micro ATX, mini-ITX, E-ATX and XL-ATX motherboard compatibility
- 4 - 5.25" bays
- 8 - 3.5" HDD trays - all compatible with SSDs
- A total of 9 expansion slots
- 3 - ModuVent™ plates – two in the top and one in the side
- 7 - Fan positions (3 fans included)
- Filtered fan slots in front and bottom
- CPU coolers up to 170 mm tall (when no fan is installed in the side panel)
- PSU compatibility: ATX PSUs up to 190 mm deep when using the bottom fan location; when not using this fan location longer PSUs (up to 345 mm deep) can be used
- Graphics card compatibility: Graphics cards up to 330mm in length with the top HDD cage installed - With the top cage removed, graphics cards up to 480mm in length may be installed
- 26 mm of space for cable routing behind the motherboard plate
- Thick rubber grommets on all holes on the motherboard plate
- Colors available: Black Pearl, Titanium Grey
- Case dimensions (WxHxD): 232 x 559 x 560mm
- Net weight: 16.4kg

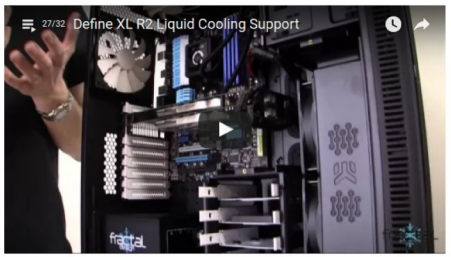
fractal-design.com

Fractal Define XL R2 Liquid Cooled System
pugetsystems.com

Solution home / Product Videos / Product Video(s) Define Series

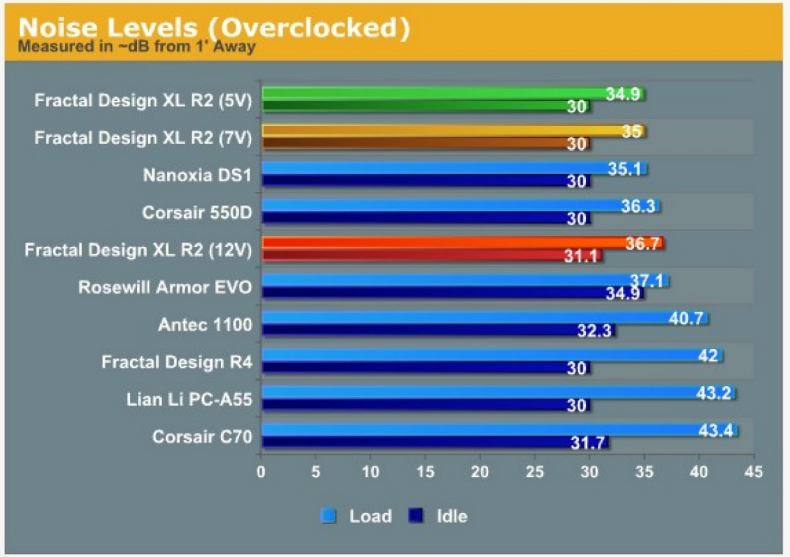
Define XL R2 Liquid Cooling

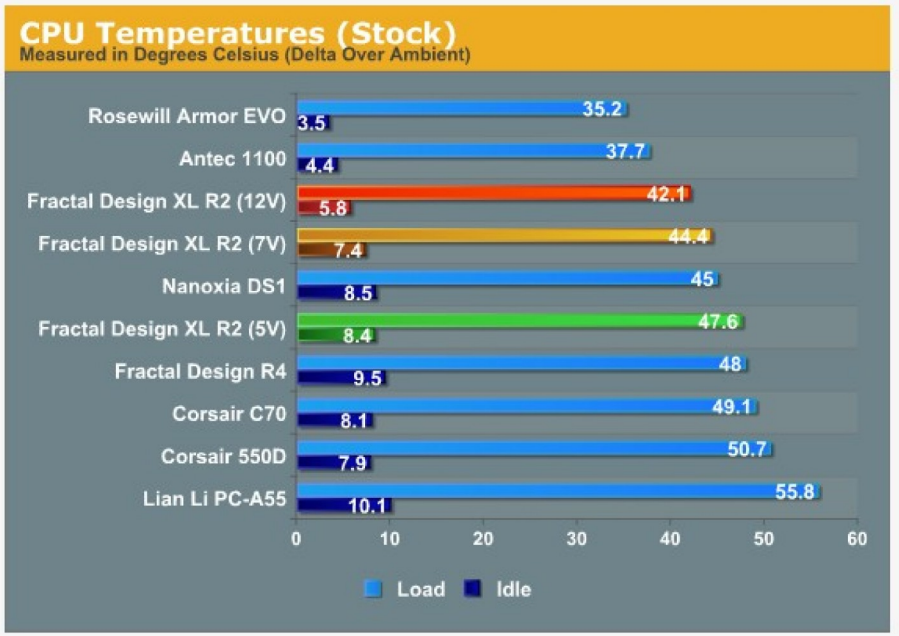
Modified on: Mon, 6 Oct, 2014 at 6:17 PM



support.fractal-design.com
www.youtube.com

The overclocked testbed usually buries quiet cases, but the XL R2 holds its own. At worst, it's competitive with Nanoxia's formerly class-leading Deep Silence 1. At best, it eclipses it soundly.





anandtech.com

Fractal Design Define XL R2

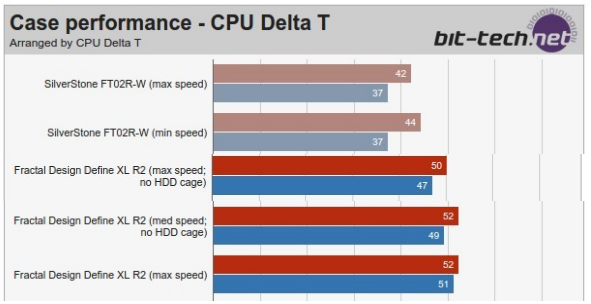
Case dimensions (WxHxD):
232 x 559 x 560mm (Fractal XL R2)
212 x 616 x 497mm (SilverStone FT02)

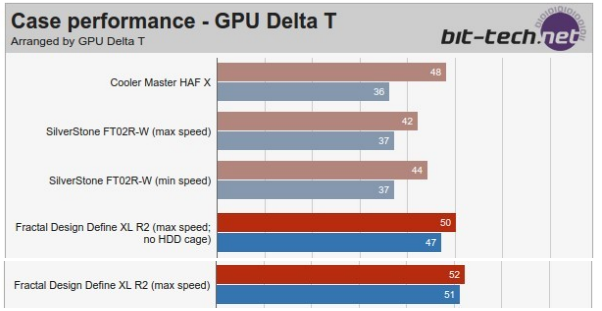
COOLING	19 / 30
FEATURES	17 / 20
DESIGN	28 / 30
VALUE	17 / 20

Score guide
Where to buy

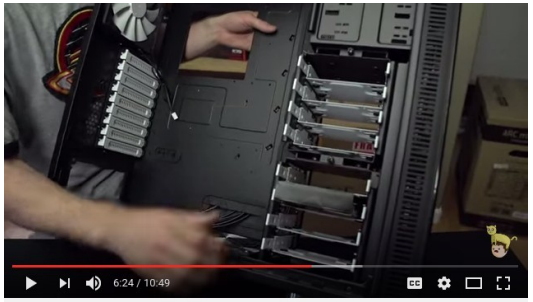
Click to enlarge -We tested at all three fan speeds with the drive cage both in and out

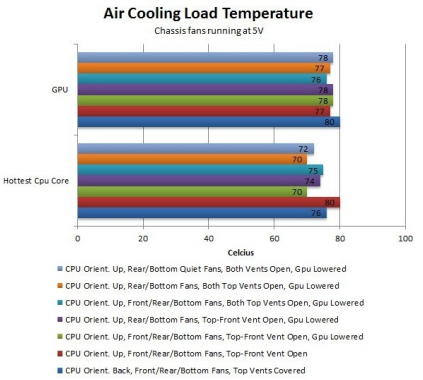
Click to enlarge - The Define XL R2 supports various hard drive cage arrangements





http://www.bit-tech.net/hardware/cases/2013/03/28/fractal-design-define-xl-r2-review/4





https://www.pugetsystems.com/labs/articles/Review-Fractal-Design-Define-XL-R2-187/page2

SilverStone FT02 SMALLER INSIDE

EXPENSIVE AND NOT THE MOST FUTURE-PROOF
GOOD COOLING PERFORMANCE

Cooling	10/10
Features	9/10
Design	9/10
Value	8/10
Overall	9/10

Score Guide

12th August 2010
by Antony Leather

SilverStone FT02R-W

Case dimensions (WxHxD):
232 x 559 x 560mm (Fractal XL R2)
212 x 616 x 497mm (SilverStone FT02)

The FT02R-W put in an **epic performance** in our tests, recording the lowest CPU delta T of any case we've tested so far of just 42°C. There isn't a lot more to it really. It wasn't quite top of the table for cooling our graphics card, managing a delta T of 37°C but this was just one degree below the Cooler Master HAF X which managed the lowest temperature. Better still, the CPU temperature only rose by 2°C and the GPU temperature actually stayed the same when we set the fans to minimum speed - a hint that the FT02R-W simply isn't harassed by our high-end test kit at all and has some cooling in reserve.

Air blasted through the chassis from bottom to top

Scan Code: LN39761
Manufacturer Code: SST-FT02B USB 3.0

In Stock

Delivered to you on Fri 12th Aug

£166.66 Ex VAT
£199.99 Inc VAT

http://www.bit-tech.net/hardware/cases/2010/08/12/silverstone-ft02r-w-review/1

Home > Computer Chassis > Fortress Series

Fortress Series FT02

- Revolutionary 90 degree motherboard mounting from RAVEN RV01
- Innovative 4.5mm aluminum unibody frame from Temjin TJ07
- Three 180mm fans for unprecedented positive pressure and stack effect cooling
- Supports liquid cooling radiator mounting
- Motherboard back plate opening behind CPU area for quick cooler assembly
- Supports 11" wide ATX motherboard
- Foam padded interior for advanced noise absorption

“Smaller” Cases FINE IF YOU DON'T NEED A LOT OF HDDS

26 mid-range desktop chassis group test

Lots of options in the mid segment of desktops

By David van Dantzig ✉ Thursday 3 October 2013 05:58

Carbide Series® Air 540

Outstanding Cooling

Our Direct Airflow Path™ design utilizes dual chambers to deliver cooler air to your CPU, graphics cards, motherboard, memory, and other PCI-E components without your drives or power supply getting in the way.

corsair.com

Popular in Deep learning builds

If comes with side window,
→ reduced sound dampening

Corsair Carbide Series Air 540

Roelof Pieters'
Deep learning build

Chillblast Fusion DEVBOX

NVIDIA DIGITS DevBox

3XS DL G10

Technical specifications

- Dual-chamber Direct Airflow Path™ design for outstanding cooling potential
- Clever, space-saving design still offers lots of internal volume
- Includes three High Performance Air Series AF140L fans for better, quieter cooling
- Tons of expansion room for high performance air cooling and liquid cooling
- Full side panel window
- Front dust filter
- Black interior
- Cable routing cutouts with rubber grommets
- CPU cutout in motherboard tray for easy CPU cooler swap-out
- Dual front USB 3.0 ports with internal connector
- Headphone, Microphone front ports
- Eight expansion slots for quad GPU installations
- Fan Mount Locations:
  - Front: 2 x 140mm (included), 3 x 120mm (pre-spaced for radiators)
  - Top: 2 x 140mm or 2 x 120mm (pre-spaced for radiators)
  - Rear: 1 x 140mm (included) or 1 x 120mm
- Dual 3.5” hot swap bays
- Four 2.5” tool-free SSD drive cages
- Maximum GPU Length 320mm
- Maximum CPU Cooler Height 170mm
- Maximum PSU length 200mm

Home > Cases/Cooling/PSUs

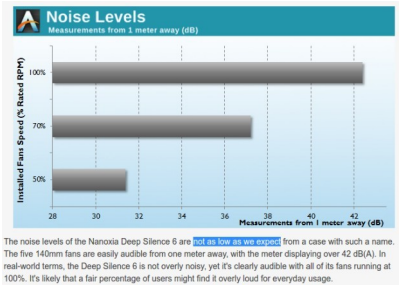
Nanoxia Deep Silence 6 Review

by E. Fylladitakis on February 13, 2014 3:00 PM EST

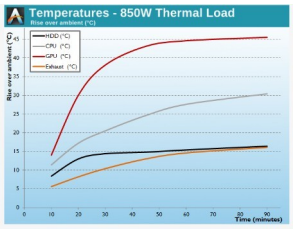
nanoxia-world.com

MODEL: NANOXIA DEEP SILENCE 6 REV. B ANTHRACITE

- Form Factor: HPTX, E-ATX, XL-ATX, ATX, Micro-ATX, Mini-ITX
- 2.5/3.5 inch drive bay internal: 10 x (max. 13) BIG! Availability issues
- Maximum installation height of CPU coolers: 200 mm
- Maximum VGA Card Length: 400 (370) mm



The noise levels of the Nanoxia Deep Silence 6 are not as low as we expec from a case with such a name. The five 140mm fans are easily audible from one meter away, with the meter displaying over 42 dB(A). In real-world terms, the Deep Silence 6 is not overly noisy, yet it's clearly audible with all of its fans running at 100%. It's likely that a fair percentage of users might find it overly loud for everyday usage.



As expected, the thermal performance of the Nanoxia Deep Silence 6 is exceptional. Even with a thermal load of 850W, the temperature readings stabilize relatively quickly, indicating that the stock cooling options of the case can cope with massive thermal loads without the temperatures rising too far above ambient. The space, multitude of ventilation openings, and ample airflow of the stock 140mm fans are sufficient for a continuous thermal load of 850W over 90 minutes.

Carbide Series® Quiet 600Q

PSU and 5.25" bay cover

Clean up the inside of your case by tucking all those cables and less-attractive drives behind a clean, refined PSU and 5.25" bay cover. Or remove them for assembly – it's up to you.

Corsair Carbide Series Quiet 600Q

Watercooling ready

Fit up to a 280mm radiator up front and up to a 360mm radiator on bottom – along with the 140mm rear fan mount, that means your next build can be cool and beautiful.

Technical specifications

- Dimensions (L x W x H): 454 x 260 x 535mm
- Maximum GPU length: 370mm
- Maximum CPU cooler Height: 200mm
- Maximum PSU length: 210mm
- Expansion Slots: 8
- 5.25" Drive bays: 2
- 3.5" Hard Drive Bays: 2
- 2.5" Drive Bays: 3
- Fans included: (x3) AF140L noise optimized 140mm fans

http://www.corsair.com/en/carbide-series-quiet-600q-inverse-atx-full-tower-case

Case security LOCKABLE CABINET

- If you need extra physical protection to your computer (shared workspace or something)
  - More options if you go for the rack-mounted server
  - Cooling might become an issue

Secure Computer Locker

This secure computer or server enclosure features a lockable computer cabinet available in small, large, and extra-large. Providing a protective shield against damage, this locking computer security case also acts as an anti-theft device for your hardware. Our locker boasts a solid 16 gauge powder coated steel construction to ensure the highest quality standard in the industry.

$199.00 SKU: 504400

https://www.tryten.com/secure-computer-locker.html

12U, 19 inch Rack with steel perforated doors, low height for standing under desks 600x800x634mm (WxDxH)

Be the first to review this product

Quick Overview

SKU: 19-6812PP

£243.00

AD server Cabinet with low-profile

19 "server rack with grid-lockable doors with door handle, variable stop
eg Ideal for installation under desks

Usable Interior Depth - 600mm

Available in six sizes

The CPU Locker Series provides computer and server protection. These Computer Enclosures protect your hardware, along with limiting access to USB drives, plugs and on/off functions.

These units work as a Server Security Enclosure, Computer Security Enclosures, or Tower Enclosure, making sure your IT equipment remains right where you left it.

Select Which Case fits your Computer

Maximum Size of Computer and Wires*							
	CPU1	CPU2	CPU2.5	CPU3	CPU3.5	CPU4	CUSTOM
Width	5"	9"	9"	10.5"	15"	21"	Request a Quote
Height	13.5"	19.5"	19.5"	25"	25"	25"	
Depth*	14"	20"	30"	27.5"	32"	27.5"	

http://www.computersecurity.com/lockdown/cpulocker.htm

8u 10" Mini Office Cabinet - 350mm Deep

Compact 10" Data Cabinet, Ideal for Small Office or Home Office Installations.

Reference: RR-W3-8-P

Availability: In Stock

The compact design of the Racky Rax 8u Cabinet makes it the ideal housing for small network requirements, allowing ... Read More

£84.00 £100.80 inc. VAT

Quantity 1

Colour Black

Add to Cart

Checkout

Images are for illustrative purposes only.

netstoredirect.com

12U, 19 inch Server Rack cabinet with glass door, low profile - for standing under desks 600x800x634mm (WxDxH)

Quick Overview

AS server cabinet with low-profile

19 "server rack with safety glass viewing door, lockable with door handle, variable stop
eg Ideal for installation under desks

Usable interior depth - 600mm

~~£242.00~~
Special Price
£217.80

Add to Cart

Add to Wishlist

Add to Compare

19power.co.uk/

PSU DETAILS IN PRACTICE

- You need to check that the **maximum power output** is sufficient for your chosen components, with the GPUs being the most power-hungry:
  - 4 x 250W GPU + 2 x 85W Dual-CPU = **1,170 W** already under full load nominally, Add 100-300W for power spikes.
- Second thing is to check whether there are connectors for all of your GPUs:
  - Easier also if you have single cable with all the GPU-side connectors (6+2 + 6 pin cable)

Titan X
8pin+6pin

GTX1080
8pin

- Third thing is the power efficiency which comes in different classes.

	DC Output	AC Input	Efficiency	Watts Lost
PSU that meets minimum ATX efficiency requirements	250W	416W	60%	166W
Typical low cost PSU	250W	357W	70%	107W
80 PLUS PSU	250W	312.5W	80%	62.5W
80 PLUS Bronze PSU	250W	294W	85%	44W
80 PLUS Silver PSU	250W	284W	88%	34W
80 PLUS Gold PSU	250W	278W	90%	28W
80 PLUS Platinum	250W	272W	92%	22W

	20% load	50% load	100% load
80 PLUS	80%	80%	80%
80 PLUS Bronze	82%	85%	82%
80 PLUS Silver	85%	88%	85%
80 PLUS Gold	87%	90%	87%
80 PLUS Platinum	90%	92%	89%

extremetech.com

UK Electricity Price 8,760 HOURS IN A YEAR 1.3 kW high estimate
0.09806 £/kWh (+20% VAT)

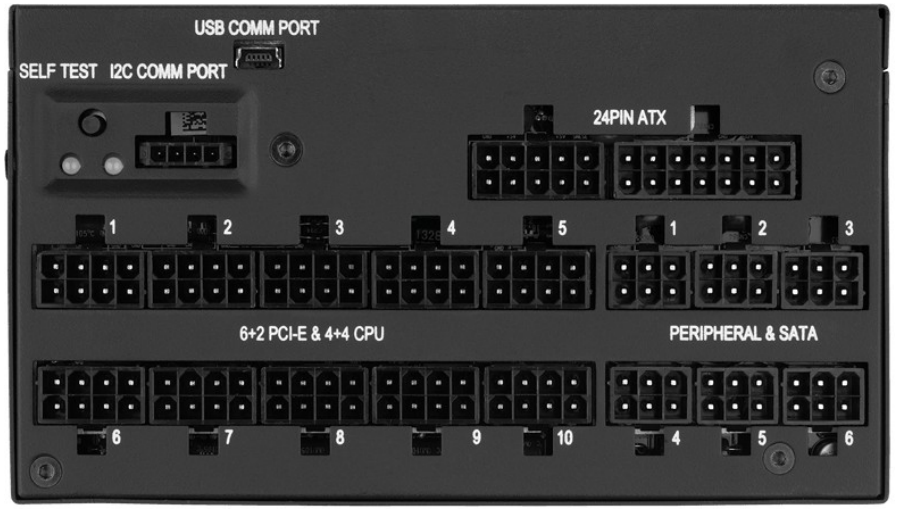
11,388 kWh at full ~1.3 kW load ➞ **£1,116** a year (100% efficiency)
12,114 kWh at full ~1.3 kW load ➞ **£1,188** a year (94% efficiency), 80 Plus Titanium
12,378 kWh at full ~1.3 kW load ➞ **£1,213** a year (92% efficiency), 80 Plus Platinum
13,090 kWh at full ~1.3 kW load ➞ **£1,288** a year (87% efficiency), 80 Plus Gold
14,235 kWh at full ~1.3 kW load ➞ **£1,395** a year (80% efficiency), 80 Plus
18,980 kWh at full ~1.3 kW load ➞ **£1,861** a year (60% efficiency)

So if you run the machine all the time at full load, you will save **£183** per year (excl. VAT) in your electricity bill when going from 80 Plus to **80 Plus Platinum**. Most likely you won't be running all the time so the saving will be less.

PSU (POWER SUPPLY UNIT) 250W PER TITAN X

- 4 x 250W (Titan X) + 2 x 85 W for CPUs

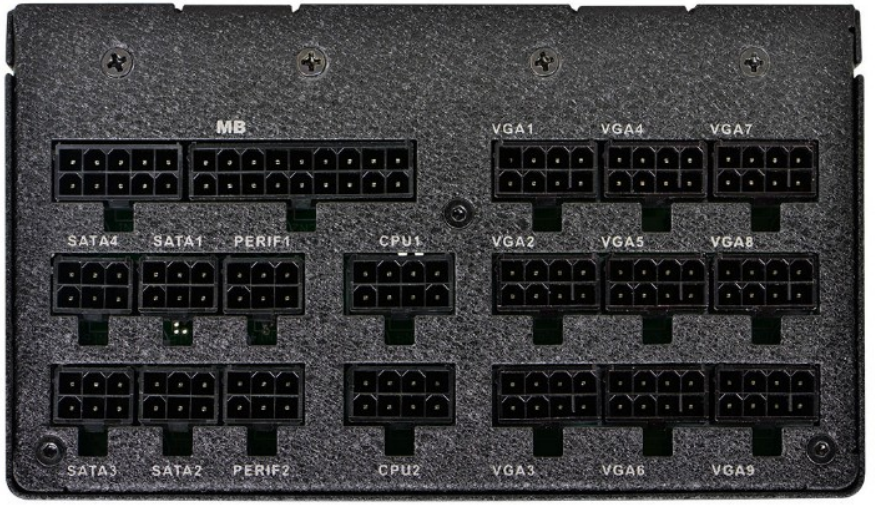
corsair.com, 1500W , £365

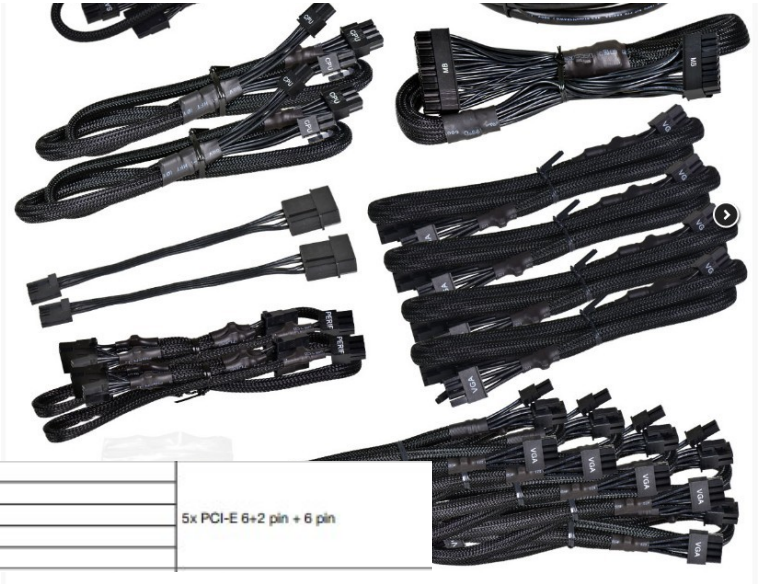


Qty	Length	Connector/cable	
4	650mm ± 10mm	1	PCI-E cable 8PIN (6+2)
2	800mm ± 10mm	1	PCI-E cable 8PIN (6+2)
2	800mm ± 10mm	2	PCI-E cable 8PIN (6+2)

2 cables per each GPU? → messier results inside your case

evga, 1600W, For £239.99, 80 PLUS Gold
£320, 80 Plus Platinum
£370, 80 Plus Titanium 94%





VGA5	5x PCI-E 6+2 pin + 6 pin
VGA6	
VGA7	
VGA8	
VGA9	

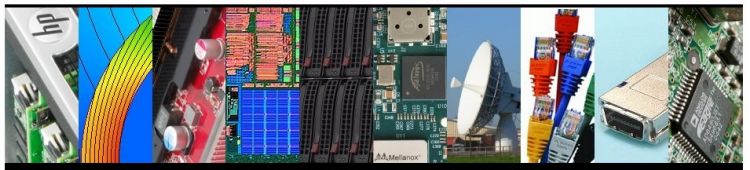
6+2 pin + 6 pin (8+6) for Titan X

UPS UNINTERRUPTIBLE POWER SUPPLY

- You would not want to have your computations stopped just because there are short power shortages.

ClusterDesign.org

Design and architecture of HPC cluster supercomputers, fat-tree and torus networks, performance modelling and other related things



Home | SADDLE | Download | Fat-Trees | Torus Networks | **UPS Sizing** | Performance Models | Floorplanning | About

UPS Sizing

Search

I am brave and bold, get me right to the tool!

Every decent supercomputer installation needs a reliable and battery-backed power supply. You need to choose your uninterruptible power supply (UPS) system based on the following requirements:

Recent Posts:
- The Journey Ends Here
- Cluster Design Tools ver. 0.8.5 — Final!
- Cluster Design Tools Updated (ver. 0.8.4), SADDLE Included
- Will We Ever See InfiniBand in Desktop Computers?

1. Power rating of the equipment that will be connected to the UPS system. In many cases, it is beneficial to provide backup power for your cooling equipment, too: in smaller machine rooms, if the cooling system goes down but computing equipment continues to operate, air temperature will immediately soar — which threatens a thermal shock to computing equipment.
2. Backup time. There must be enough time to automatically save work (or create restore points), and then cleanly shut down applications and operating systems on compute nodes. Alternatively, if there is a backup diesel generator, the UPS backup time must be large enough for the diesel generator to start and reach its steady-state regime.
3. Physical constraints, such as size and weight of UPS equipment. Sometimes you only have a limited number of empty units in a rack, or space for new racks on the floor.
4. Budgetary constraints.

http://clusterdesign.org/ups-sizing/

Problem: High-wattage UPS is quite expensive

1,980W

APC SMT2200RMI2U Smart-UPS,1980 Watts /2200 VA,Input 230V /Output 230V, Interface Port USB, Rack Height 2 U

£880

APC Back-UPS Pro 1500 - UPS - 865 Watt - 1500 VA(BR1500G)

£200

865W

SURGE PROTECTOR: Cheap, offer some protection from spikes

Our pick

Cuts off power if it can't stop surges

Tripp Lite TLP1008TEL

This 10-outlet surge protector stops power delivery once the protection wears out while also meeting our bar for quality.

Buy from Amazon

*At the time of publishing, the price was $32.

Tripp Lite TLP1008TEL
thewirecutter.com

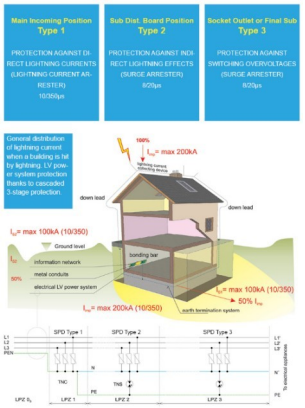
Does not work if surge components die! See below for details

OPINION

A surge protector that doesn't protect

Computerworld | Jan 25, 2012 4:25 PM PT

computerworld.com



Keyboard & Mouse

Inexpensive way to increase productivity with ergonomic interface

- subjective what is good, go and try and convince your boss to pay for it :P

Our pick

The Microsoft Sculpt Ergo meets all our ergonomic criteria, making it best for most people.

The Microsoft Sculpt Ergo is the only keyboard we tested that offers both tenting—rotating the wrists properly to avoid ulnar deviation—and a negative tilt to prevent extension. The manta-ray-shaped keyboard is designed with a curved bump in the middle to achieve tenting of about 10 degrees, while a magnetic attachment tilts the back of the keyboard down about 5 degrees. (Yes, I got out my protractor for this.)

Our pick

$68 from Amazon

Affordable and comfortable

Microsoft Sculpt Ergo

An inexpensive ergonomic keyboard that puts your wrists in the ideal typing position for pleasant typing over long periods.

http://thewirecutter.com/reviews/comfortable-ergo-keyboard/

Our pick

Logitech Marathon M705

The Logitech Marathon M705 was universally liked by our panel of experts and laypeople. It fits a variety of hand sizes, and it grips and tracks accurately on everything from desks to wood floors to fabric—just not on glass or mirrors.

Buy from Amazon

*At the time of publishing, the price was $34.

http://thewirecutter.com/reviews/best-wireless-mouse/

PCMag UK | First Looks | Logitech MX Master | Review

Logitech MX Master

EDITOR RATING: EXCELLENT (4.0) APR 04, 2015

PROS

Elegant design. Multiple connectivity options. Can switch easily for use among multiple devices. Fast-charging battery, with long life on our tests. USB cable included.

CONS

Touchy adaptive scrolling. Gesture button and motions are unintuitive.

BOTTOM LINE

The Logitech MX Master is a wireless computer mouse that supports up to three different devices, but some features still need refinement.

uk.pcmag.com | Logitech MX Master

http://www.trustedreviews.com/mx-master-mouse-review

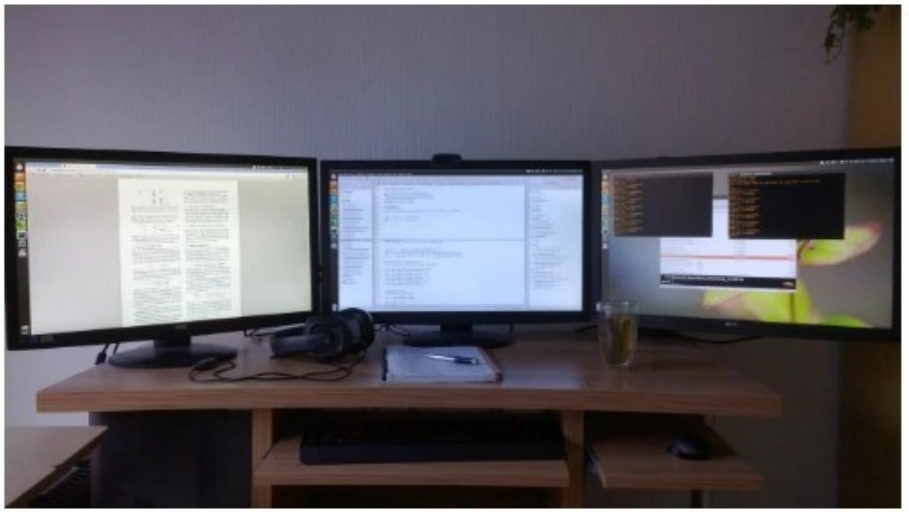
Monitor | PERSONAL PREFERENCES

http://timdettmers.com/2014/09/21/how-to-build-and-use-a-multi-gpu-system-for-deep-learning/

Monitors

I first thought it would be silly to write about monitors also, but they make such a huge difference and are so important that I just have to write about them.

The money I spent on my 3 27 inch monitors is probably the best money I have ever spent. Productivity goes up by a lot when using multiple monitors. I feel desperately crippled if I have to work with a single monitor. Do not short-change yourself on this matter. What good is a fast deep learning system if you are not able to operate it in an efficient manner?



Typical monitor layout when I do deep learning: Left: Papers, Google searches, gmail, stackoverflow; middle: Code; right: Output windows, R, folders, systems monitors, GPU monitors, to-do list, and other small applications.

If you work with images, makes sense to have a monitor that has somewhat *accurate color reproduction*, and *good gamut*. In other words, avoid tn-panels, and look for monitors that are good for photo/video editing (see e.g. GUIDE from pc4u.org)

13 best budget displays 2016 UK: what's the best budget PC monitor?

Top 13 sub-£200 flat-panel monitors you can buy in the UK in 2016

1. BenQ RL2460HT

- **Rating:** ★★★★½
- **Reviewed on:** 1 March 16
- **RRP:** £168.11 inc VAT
- **Buy** from Amazon for £164.40

The BenQ RL2460HT is great gaming monitor which also doubles as a fantastic monitor for editing photos or videos, since it has good gamut and accurate colour reproduction. The HDMI output makes it a great choice for those who take gaming seriously and want to record their exploits.

RL2460HT is a **24" 1920x1080 TN display** with an **LED backlight**
"...great colour accuracy really sets apart the RL2460HT from other TN displays"

£187.49 Inc VAT
BenQ BL2420PT

Asus VX24AH

Asus PB258Q

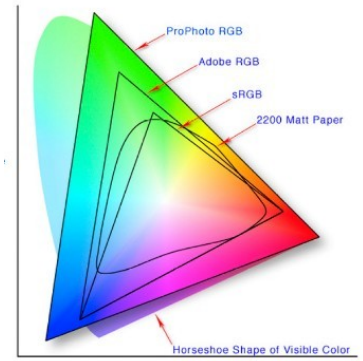
Monitor DETAILS, FROM PC4U.ORG

What Type of Screen Should You Get?

In terms of pricing, the cheapest monitors are equipped with twisted pneumatic **(TN) displays**. While these monitors have the fastest response time, unfortunately, they are limited when it comes to color reproduction, limited narrow viewing angles (limited to 160°), and poor black levels. Because of these limitations, these monitors are ideal for gaming and basic word processing. Therefore, if you're looking to purchase a new monitor for serious photographic work, I would not recommend getting a TN monitor.

Any serious photographer should be looking to purchase at least an **IPS screen**. Although, IPS screens used to be the most expensive monitors on the market, the prices have considerably dropped in the last few years and have become more affordable. Despite a slower response time (not necessary for photo editing), the IPS screen allows for much deeper blacks, better color renderings, balanced and improved contrasts, and viewing angles of almost 180°.

VA – (Vertical Alignment) monitors are LCDs with vertical alignment and represent the middle between TN and IPS. VA monitors allow for good viewing angles and color renderings that are superior to TN's although the response time is slower. This could lead to one disadvantage of resulting in a very high contrast value.

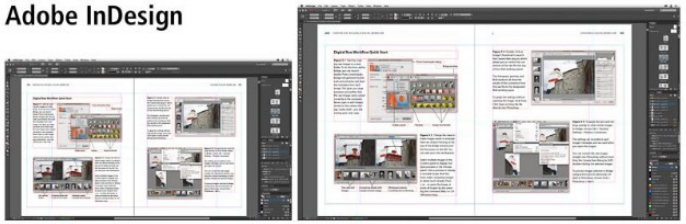


Most displays can reproduce a **gamut** found in your standard sRGB (red, green, blue), which is a standard RGB space created by Microsoft and Hewlett Packard in 1998. Some displays offer a wide gamut that can duplicate as large as the Adobe RGB gamut. Usually, this gamut only benefits those people who perform color-critical work that are often reproduced on CMYK presses or fine-art printers, requiring a much larger color gamut.

Adobe InDesign

16:9 (1920 x 1080 px) **16:10** (1920 x 1200 px)

Adobe InDesign



1920 x 1080 px 2560 x 1440 px

Check resolution:
Nicer (subjective opinion) to have more content fitting to screen

Glossy or Matte?

Which one is better? Glossy appearances favor vibrant and contrasting colors, which are suitable for computers primarily used for entertainment purposes. Unfortunately, with photographic work, glossy screens introduce a number of disadvantages.

Glossy vs. Matte Monitor

Although shiny, glossy screens may look extremely attractive and beautiful at first glance, the reflections that are caused by them will lead you to have second thoughts.

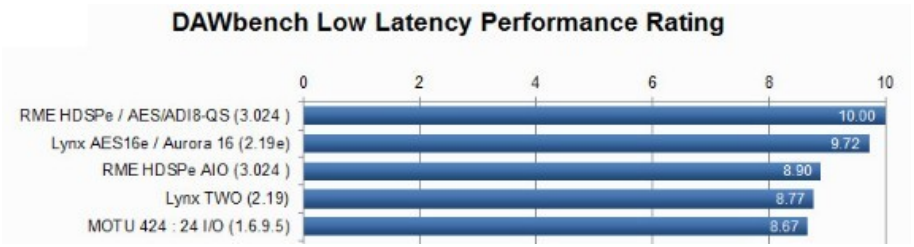
Other Peripherals

OPTICAL DRIVE

All cases pretty much can house at least one **optical drive** (5.25" slot; 4 slots in Fractal Define XL R2), or use external drive with USB connection.

SOUND CARD

Go for low-latency USB card if you do auditory psychophysics (or music production) with the same machine, otherwise the integrated sound card probably is more than sufficient for you.



https://www.gearslutz.com/board/11591001-post771.html
e.g. RME Babyface, £550; M-Audio MTrack, £70;
Behringer UCA222, £25

CASE SILENCING TRY DYNAMAT?

CASE FANS

To improve cooling, you may want to add silent **case fans** with PWM pin so that they can be automatically adjusted by motherboard.

To differentiate between standard 3-Pin fans and 4-pin PWM (self adjusting) fans we use PWM in the name of the fan. If it doesn't say PWM in the name of the fan then it is a standard 3-pin fan.

To convert airflow in cubic feet per minute (CFM) to cubic metres per hour (m³/h), multiply the CFM figure by 1.7.

Name	Price	Airflow (CFM)	Noise (dBA) ▲	Max speed (RPM)
Noctua NF-P14s REDUX 900	£12.40	49.2	13.2	900
Nanoxia DS 140mm Fan	£10.54	68.5	14.4	1100
Nanoxia DS 140mm PWM Fan	£10.75	76.5	16.2	1400
Noctua NF-A14 ULN	£17.20	67.9	19.2	1200
Noctua NF-A15 PWM	£17.20	67.9	19.2	1200
Noctua NF-A14 FLX	£17.20	67.9	19.2	1200

https://www.quietpc.com/140mmfans
e.g. Nanoxia NDS Deep Silence Fan 140 PWM, ~£15

FAN HUBS

You can use several case fans with the same PWM control signal, using a **fan hub** if you prefer your fans to be silent with no load and start operating when training your network.

Phobya PWM 4x 4Pin Splitter
kustompcs.co.uk, £5

4-Pin PWM Power Distribution PCB 4-Way Block Fan Hub Power Splitter

Office Thermal Management

- In addition to the cost of the electricity, you should take into account that your deep learning **computer** will now essentially works **as a heater**.
  - Imagine being in a small office with your desktop with no air conditioning. During winter time, you then save in heating.
  - Harder to fix if you are stuck in an office with no AC

Air Natural Anna Little Space Heater, ANNA0002
1200 wattsW
by STADLERFORM
8 customer reviews | 3 answered questions

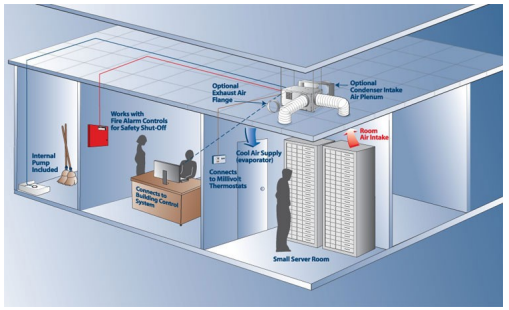
RRP: £59.00
Deal Price: £49.00 & **FREE Delivery** in the UK. Details
You Save: £10.00 (17%)

In stock. **1,200W as well**

Want it tomorrow, 9 Aug.? Order it within **18 hrs 54 mins** and choose **One-Day Delivery** at checkout. Details

Dispatched from and sold by Amazon. Gift-wrap available.

3 new from £49.00 1 used from £50.00



The Movincool CM12 Ceiling-Mounted Air Conditioner

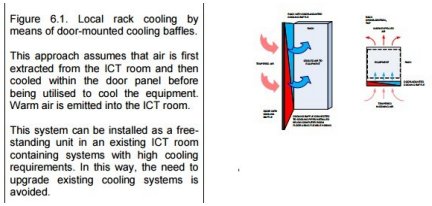


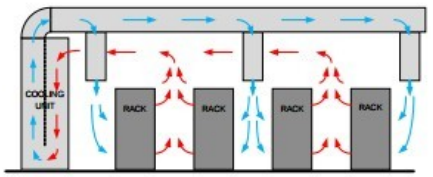
Figure 6.1. Local rack cooling by means of door-mounted cooling baffles.

This approach assumes that air is first extracted from the ICT room and then cooled within the door panel before being utilised to cool the equipment. Warm air is emitted into the ICT room.

This system can be installed as a free-standing unit in an existing ICT room containing systems with high cooling requirements. In this way, the need to upgrade existing cooling systems is avoided.

Figure 5.3. Partially regulated input air in a room without a raised computer room floor.

Maximum cooling capacity of 3 kW per rack. Inexpensive, easy to install.



Ventilation and Cooling Requirements for ICT rooms services.geant.net

Server Furnace AS BUSINESS

- With the InfiniBand (40, 50 Gb/s) or similar fast local network, you could actually in theory do the heating of your offices if you place for example one node in one room.

ars TECHNICA UK BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE FORUMS

MINISTRY OF INNOVATION —

Data furnaces arrive in Europe: Free heating, if you have fibre Internet

Nerdalize is rolling out eRadiators in the Netherlands, providing 1000W of heat.

SEBASTIAN ANTHONY - 28/5/2015, 12:00

The Nerdalize eRadiator. (The real thing doesn't have the logo or tacky slogan on the side, thankfully.)

Back in 2011, Microsoft Research published a research paper on the topic of **data furnaces**. The concept was simple. Microsoft has a lot of servers, mostly sitting in large data centres, producing huge amounts of heat—heat that is a massive nuisance to deal with. Instead of venting that heat into the environment (and spending a fortune in the process), why not do something useful with it?

Nerdalize is a small Dutch company that is trying to commercialise Microsoft's data furnace idea. The first product is the eRadiator, which, given its size, probably contains **two or three servers that pump out around 1000W of heat**—probably just enough to heat a small room in winter. In an interview with the BBC, one of the first users of the eRadiator says it takes "about an hour" to heat up.

In exchange for **free heating** (after the €400-500 setup cost), Nerdalize uses the network of eRadiators to provide a cloud computing service. Because the company doesn't run a centralised data centre, operating costs are much lower, which means the **"cost-per-job [to the customer] is up to 55% lower**." The quality-of-service will be be lower than centralised cloud compute, too—Nerdalize won't have any control over the access network (what if the home owner decides to do some torrenting?)—but there are plenty of use cases where cost is more important than latency.

http://arstechnica.co.uk/business/2015/05/data-furnaces-arrive-in-europe-free-heating-if-you-have-fibre-internet/
Sebastian Anthony - 28/5/2015

Summary

- **What did we learn?** Not that difficult to pick the components for your workstation → you will save money and understand your own system better

- Now we can reverse-engineer the costs of commercial build, and you can yourself determine if you get the 'markup value' from buying a COTS (Commercial off-the-shelf computer).

Reverse-Engineer COTS #1

WhisperStation™- Deep Learning

- Intel Xeon E5-1600v4 CPU, 6-core
- Quad-Channel 2400 MHz / 2133 MHz ECC/Registered Memory (8 slots), 64 GB
- Four NVIDIA GeForce GTX Titan X "Pascal" GPUs
- 250 GB SSD for the operating system
- 6TB of high-speed RAID10 storage
- Motherboard?
- High-efficiency (80 PLUS certified) quiet power-supply
- Case?
- CPU Fan
- Case Fan

= $10,956+

= £8,480+

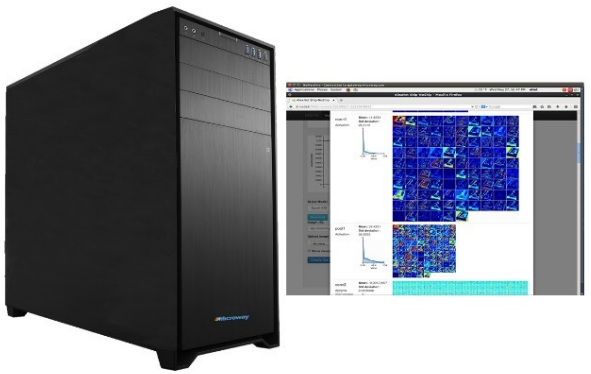
- £275 (E5-1620), £500 (E5-1650) single-CPU Xeon System
- £370 2400 MHz / £300 (2133 MHz) ECC/
- £3,720 (Pascal, $1,200 each)
- £110 (Samsung 850)
- £600, 4 x 3TB, 7200 rpm
- £400 AsusX 99
- £270, 80 Plus Gold
- £100, ballbark guess
- £50
- £60

= £6,110

Now missing the following components from BoM:
Monitor, **~£200**
Keyboard, **~£60**
Mouse, **~£70**

Supported for Life

Our technicians and sales staff consistently ensure that your entire experience with Microway is handled promptly, creatively, and professionally.



WhisperStation™- Deep Learning

Ultra-Quiet Computing for Deep Learning Researchers

Reverse-Engineer COTS #2

3XS Deep Learning G10

LN57697 3 x 120mm Noctua NF-F12 Case Fan	£60
LN74166 4 x 12GB NVIDIA TITAN X	£2,680, (Maxwell) | £3,720 (Pascal, $1,200 each)
LN62019 1600W EVGA SuperNOVA, 80PLUS Titanium 94%	£370
LN58578 256GB Samsung 850 PRO, 2.5" SSD, SATA III	£124
LN53870 3 x 3TB WD3003FZEX Black, Performance 3.5" HDD, SATA III, 7200 rpm	£450
LN61212 3XS Only Corsair SP120 PWM Quiet Single Fan (Red/Blue/White Ring)	£16
LN64278 512GB Samsung SM951, M.2	£225
LN61156 64GB (8x8GB) Corsair DDR4 Vengeance PC4-19200, 2400 MHz	£310
LN60298 Asus X99-E WS, Intel X99	£470
LN51461 Corsair Carbide Series Air 540 Case	£130
LN72520 Corsair Hydro Series H45, 120mm All-In-One Hydro Cooler CPU Cooler	£40
LN72344 Intel Core i7 6900K	£950
LN58663 StarTech.com 3 Drive 3.5in SATA/SAS Hot Swap Hard Drive Backplane	£66

£8164.84 inc VAT

=£5,890 inc VAT (Maxwell), markup of **~£2,275**

=£6,930 inc VAT (Pascal), markup of **~£1,235**

The markup now includes

LN64660 *3XS LOGO* 16GB Kingston USB3 Pendrive for Windows Recovery Media
LN61909 6 x 3XS System Build
LN71540 3XS Systems 7 Day Technical Support
LN45708 Scan 3XS System - 3 Year Warranty (Mainland UK) - 1st Year Onsite, 2nd & 3rd Year Return to Base (Parts & Labour)

Now missing the following components from BoM:
Monitor, **~£200**
Keyboard, **~£60**
Mouse, **~£70**

Reverse-Engineer COTS #3

DIGITS DevBox includes:

- Four TITAN X GPUs with 12GB of memory per GPU
- 64GB DDR4, 2166 MHz
- Asus X99-E WS workstation class motherboard
- Core i7-5930K 6 Core 3.5GHz desktop processor
- Three 3TB SATA 6Gb 3.5" Enterprise Hard Drive in RAID5
- 512GB PCI-E M.2 SSD cache for RAID
- 250GB SATA 6Gb Internal SSD
- 1600W Power Supply Unit from premium suppliers including EVGA
- Corsair Carbide Air 540 Case
- Pre-installed standard Ubuntu 14.04 w/ Caffe, Torch, Theano, BIDMach, cuDNN v2, and CUDA 7.0

= $15,000 (?)

- **= ~ £11,600**

Approx. prices (August 2016)

- £2,680
- £250
- £400
- £530
- £405, 7,200rpm
- £280
- £110
- £270, 80 Plus Gold
- £100
- Ubuntu free
  like all the libraries as well

= £5,025

Now missing the following components from BoM:

- Monitor, **~£200**
- Keyboard, **~£60**
- Mouse, **~£70**

= £5,355

- CPU Cooler (if not boxed, like i7?), **~£50**
- Additional fans?, **~£4x15**

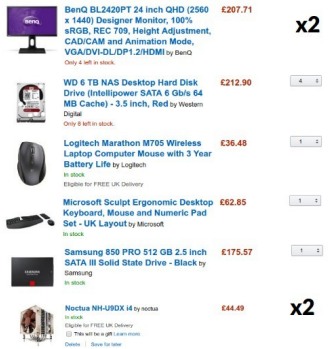
= £5,465

Cap the system cost to ~£11,600

THE COST OF NVIDIA DIGITS DEVBOX

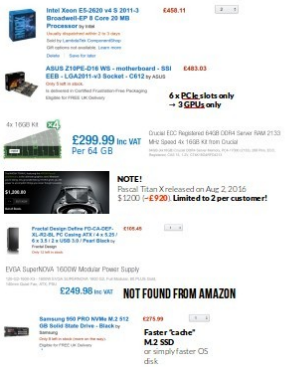
We can now use the Supermicro MB + Chassis as basis see previous slides for further details

"BASE"



~ £1,600

"INTEL DUAL-XEON CONSUMER"



Intel Xeon BASE Price
~ £1600 previous slide

+£920	Xeon-specific
+£480	~£2,600
+£1,200	
+(3x£920)	Xeon and i7
+£100	~£3,400
+£250	
+£275	

~ £7,600

"SUPERMICRO"

Supermicro X10DRG-Q

~(£1,150 - 850), thus **£300 more expensive** build than the "consumer dual XEON options"

~£390

~ £7,600 + £300
= ~£7.900

~£7,900 + £920 =
~£8,800 4* Titan-X

~9.000 + £170 (RAID CACHE) **= ~£9,200**

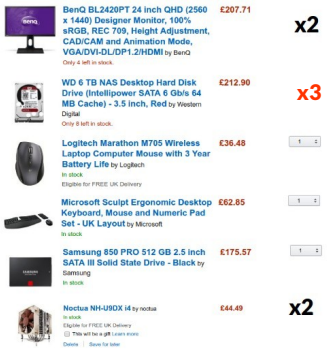
We see now that the **CONSUMER DEAL** is **NOT VERY GOOD** if you want to go for more memory with Intel Xeon systems. With only £300 more you can upgrade the consumer motherboard to the Supermicro server solution (MB + chassis&power) and you can throw even more more NVIDIA Titan X to the build giving you 44 TFLOPS (FP32) of GPU crunching power.

Supermicro solution would leave us still "2x PCI-E 3.0 x8 (1 in x16)", and 1x PCI-E 2.0 x4 (in x8) to be used for **Infiniband Network cards** (40Gb/s) if you want to build GPU clusters (see coming slides) using NVIDIA GPUDIrect RDMA. We now have over **£2,000** of "extra money" to spend if you would like to match the price of NVIDIA Digits Devbox.

Disadvantage of the Supermicro solution is that it does not support the latest **SSD drives with M.2** interface, get another SATA III SSD for the RAID cache (add ~£170).

“Budget” Supermicro Server

- Start with the bare minimum and add more over time:



x2

x3

x2

~£1,400

Western Digital Red Pro

Prices (Aug 2016), 7200 rpm

2 TB	£120	£60/TB
3 TB	£135	£45/TB
4 TB	£199	£50/TB
6 TB	**£213**	**£36/TB**
8 TB	£430	£54/TB

RAID CACHE (~£175)

Samsung 850 Pro SSD

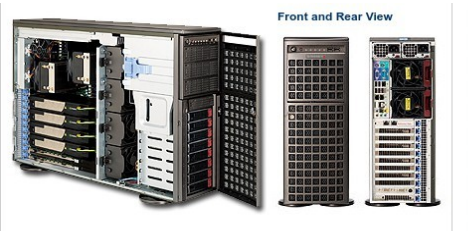
Prices (Aug 2016), Sata III

1024 GB = 1 TB

120GB	£70	£597/TB
250GB	£110	£451/TB
500GB	£175	£358/TB
1 TB	**£337**	**£337/TB**
2 TB	£734	£367/TB

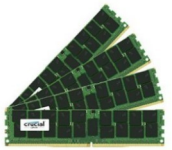
2x Pascal Titan X GPU

~£1840



SC747TQ-R1620B chassis to house the X10DRG-Q motherboard

$1,500 - **£1,160**



Crucial CT4K16G4RFD424A DDR4, 64 GB (4 x 16 GB), DIMM, 288-Pin, 2400 MHz, PC4-19200, CL17, 1.2 V Internal Memory

by Crucial

Be the first to review this item

RRP: £762.00

Price: £320.50 FREE UK delivery.

You Save: £441.50 (58%)

Only 2 left in stock - order soon.

Estimated delivery 16 - 18 Aug. when you choose Express Delivery at checkout.

Details

8 x RDIMMs

128 GB

~£640

Note!

For quad-channel setup with Dual-CPU, you need 8 slots occupied. If you really want 64 GB you would need to buy 8 x 8 GB RDIMMs

Intel Xeon E5-2620 v4 S 2011-3 Broadwell-EP 8 Core 20 MB Processor by Intel

£458.11

2

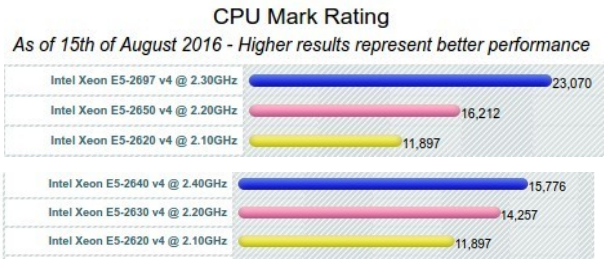
Usually dispatched within 2 to 3 days

Sold by LambdaTek ComponentShop

Gift options not available. Learn more

Delete | Save for later

~£900



Benchmarks for single-CPU system

£5.000
£2.150
£900
£1,740
£1.230
£900

Now if you are under *tranche* funding scheme, we can start with a system cost of:

1400+175+1160+1840+640+900 ~ **£6,100**

Add 2 x Titan X (~£1,840), 128 GB memory (£640) and storage (5 x 6 TB ~ £1,080) that you want later

~£3,560 ~ £9,700

Use the Asus ESC8000 G3 as node

Upgrade from the proposed system

3x TITAN X GPUs (Pascal) with 12GB of memory per GPU	£2,790 (Pascal, $1,200 each)	
256GB (16x 16GB) DDR4, 2400 MHz	£1,280	**-£640** (if 128 GB is ok)
2x Intel Xeon E5-2620 v4	£830	
2x Noctua NH-U9DX i4	£90	
4x 6TB SATA 3.5"	£850	
512GB M.2 SSD	£275	
512GB SSD SATA3	£175	
2x BenQ BL2420PT (2560x1440 IPS)	£415	
Logitech Marathon M705 Mouse	£40	**+ £30** (to upgrade the mouse)
Microsoft Sculpt Keyboard	£65	
	= £6,800	**= £6,200**
		3 X 11 TFLOPS **= 33 TFLOPS**

Asus ESC8000 G3, ASUS Z10PG-D24 Server Board	£3,330	
HDD Enclosure for 4 x 3.5" HDDs	~£100	
	= 10,200	**= £9,600**

You can add now more GPUs:

5x TITAN X GPUs (Pascal) with 12GB of memory per GPU	£4,650	
	= £14,850	**= £14,250**
		8 X 11 TFLOPS **= 88 TFLOPS**

ASUS

https://www.asus.com/uk/Commercial-Servers-Workstations/ESC8000_G3/

Asus ESC8000 G3 (ASMB8-IKVM)
3U Rack Server

SYSTEM SPECIFICATIONS	
GPUs	8x Tesla P100
TFLOPS (GPU FP16 / CPU FP32)	170/3
GPU Memory	16 GB per GPU
CPU	Dual 20-core Intel® Xeon® E5-2698 v4 2.2 GHz
NVIDIA CUDA® Cores	28672
System Memory	512 GB 2133 MHz DDR4
Storage	4x 1.92 TB SSD RAID 0
Network	Dual 10 GbE, 4 IB EDR
Software	Ubuntu Server Linux OS DGX-1 Recommended GPU Driver
System Weight	134 lbs
System Dimensions	866 D x 444 W x 131 H (mm)
Packing Dimensions	1180 D x 730 W x 284 H (mm)
Maximum Power Requirements	3200W
Operating Temperature Range	10 - 30° C

The NVIDIA DGX-1
$129,000
£100,000

with option to add NVIDIA DGX-1 support.

170 TFLOPS (FP16), **£590 per TFLOP**
84.8 TFLOPS (FP32) **£1,180 per TFLOP**

With **~£15,000** you could get all the 8 GPU slots filled

This would give 88 TFLOPS (FP32) which would actually exceed the FP32 performance of DGX-1 (on paper). Note that some of the specs on DGX-1 are better than with our hypothetical budget version

You could build a **GPU cluster** with 7 of these 3U rack servers for ~£105,000 which would give you 7 x 88 FLOPS (FP32) = **616 TFLOPS (£170 per TFLOP, ~7.3x of DGX-1)** assuming idealistic linear scaling (which is not the case as you saw from previous slides).

~7x bang-for-the buck compared to DGX-1
IN PRACTICE WOULD BE INTEREST TO SEE SOMEONE BENCHMARKING THIS

Modified from: http://www.nvidia.com/object/deep-learning-system.html

Asus ESC8000 G3 BAREBONE MINIMUM

2x TITAN X GPUs (Pascal) with 12GB of memory per GPU	£1,940 (Pascal, $1,200 each)
128GB (8x 16GB) DDR4, 2400 MHz	£640
2x Intel Xeon E5-2620 v4	£830
2x Noctua NH-U9DX i4	£90
3x 6TB SATA 3.5”	£640
512GB M.2 SSD	£275
512GB SSD SATA3	£175
2x BenQ BL2420PT (2560x1440 IPS)	£415
Logitech Marathon M705 Mouse	£40
Microsoft Sculpt Keyboard	£65
Asus ESC8000 G3, ASUS Z10PG-D24 Server Board	£3,330
HDD Enclosure for 8 x 3.5” HDDs	£200
= £8,640	2 X 11 TFLOPS **= 22 TFLOPS**

Upfront cost now is **£2,500** more than for the ““Budget” Supermicro Server” built from Supermicro Server motherboard (space for 4 GPUs, ~£6,100)

You can add now more GPUs:

6x TITAN X GPUs (Pascal) with 12GB of memory per GPU	£4,600 (Pascal, $1,200 each)
	= £13,440 8 X 11 TFLOPS **= 88 TFLOPS**

You can add now more RAM:

256GB (16x 16GB) DDR4, 2400 MHz	£1,280
	= £14,720

More HDDs

5x 6TB SATA 3.5	£5x215 ~ £1075
	= £15,820

The **“full cost”** now is **£5,500 more** than for “Budget” Supermicro Server” built from Supermicro Server motherboard (space for 4 GPUs, ~£9,700)

With **~1.6x** investment (£15,200/£9.700) you get the **2x** of flops assuming linear scaling of performance.

supermicro.com

We have now also 24 DIMM slots compared to 16 offered by the Supermicro. So with 16 GB RDIMMs we get a total of **384 GB RAM.**
The Asus board actually support maximum of **1536 GB of LRDIMM** if the 384 GB does not seem enough for you. One 64 GB LRDIMM (2400 MHz, Crucial) costs around ~£470, thus the 1536 GB would cost around £11,300 (compared to £1,920 for 384 GB of RDIMM).
8 x 32GB of **LRDIMM** would cost **£1,920** (8 x £240) giving you quad-channel performance for both CPUs and 256 GB of LRDIMM RAM with 16 slots still free for future upgrades.

https://www.asus.com/uk/Commercial-Servers-Workstations/ESC8000_G3/specifications/

Asus ESC8000 G3 PERFORMANCE CORRECTION?

System	Hours to Train	Speed-up vs Xeon Phi
PC with 4x NVIDIA TITAN X (Maxwell) ** based on NVIDIA Caffe implementation as of March 2015*	25 Hours	-
Four Xeon Phi servers	10.5 Hours	-
PC with 4 NVIDIA TITAN X (Maxwell) ** based on publicly available Caffe as of August 2016, cuDNN5*	8.2 Hours	1.3x faster than Xeon Phi
PC with 4 NVIDIA TITAN X (Pascal) ** based on publicly available Caffe as of August 2016, cuDNN5*	5.5 Hours	1.9x faster than Xeon Phi
NVIDIA DGX-1	2 Hours	5.3x faster than Xeon Phi

https://blogs.nvidia.com/blog/2016/08/16/correcting-some-mistakes/

NOMINAL VALUES

4 X 11 TFLOPS **= 44 TFLOP/S** * (5.5 * 60 * 60 S) = 871,,200 TFLOP

8 X 10.6 TFLOPS **= 84.8 TFLOP/S** * (2 * 60 * 60 S) = 610,560 TFLOP

NVIDIA computed the task in 70% of the time needed by Titan X cluster with nominal flops. Thus the nominal should be **multiplied** (1/0.7) by **~1.43** to get "equivalent FLOPS" (note! Totally non-standard term)

So going back to previous slides:

8 x 10.6 *"DGX-1"* TFLOPS would correspond to 84.8*1.43 *"Titan X Maxwell"* FLOPS

~ 121.23 TFLOPS

Thus for new "bang-for-buck" measure we would still get **616 TFLOPS (£170 per TFLOP)** with 7 x Asus ESC8000 with each having 8 Titan X (Pascal), but the DGX-1 would be corrected up to 121.23 TFLOPS (£100,000/121.23 ~ **£825 per TFLOP**) narrowing the bang for buck to **~4.85** which still seems quite good.

Recently Baidu team (Diamos et al., 2016 "Persistent RNNs: Stashing Recurrent Weights On-Chip") showed that one gets almost linear scaling with 128 Titan X (Maxwell) GPUs lending support for the quick'n'dirty performance estimates here and the cost-effectiveness of Asus ESC8000 G3 instead of the DGX-1.

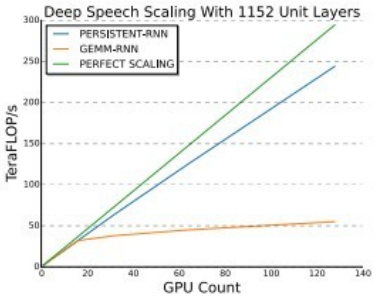


Figure 5: Throughput scaling of the 48 RNN, 61 total layer RNN with a fixed algorithmic mini-batch of 512.

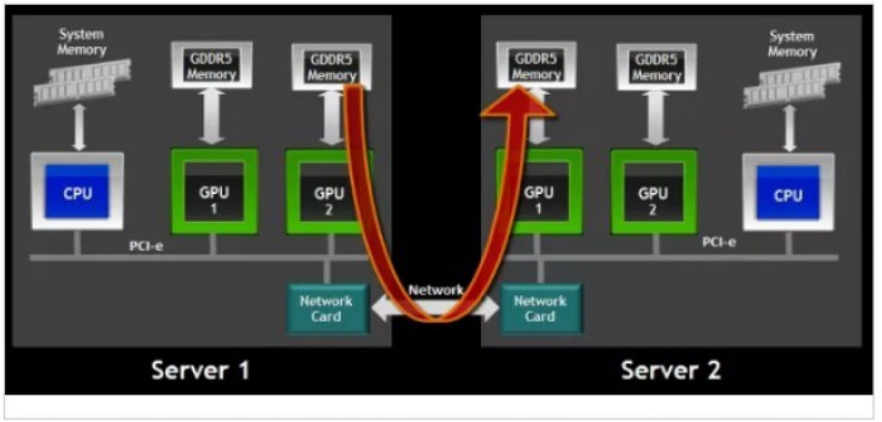
Workstation Conclusion

- Depending on your budget, and future needs but it seems that the **server-based solutions are better choice** for university/startup settings.
  - The cost of Supermicro MB based solution is quite close to "consumer" dual-CPU solution in cost and allows one more GPU to be used with extra PCI slots left for network adapter and possible non-"deep learning" GPU just for display purposes.
  - The upfront cost of the Asus ESC8000 G3 solution is higher but then scales up to 8 GPUs giving ~72.5% of the performance of the "supercomputer DGX-1" at ~16% of the cost.

GPU Cluster

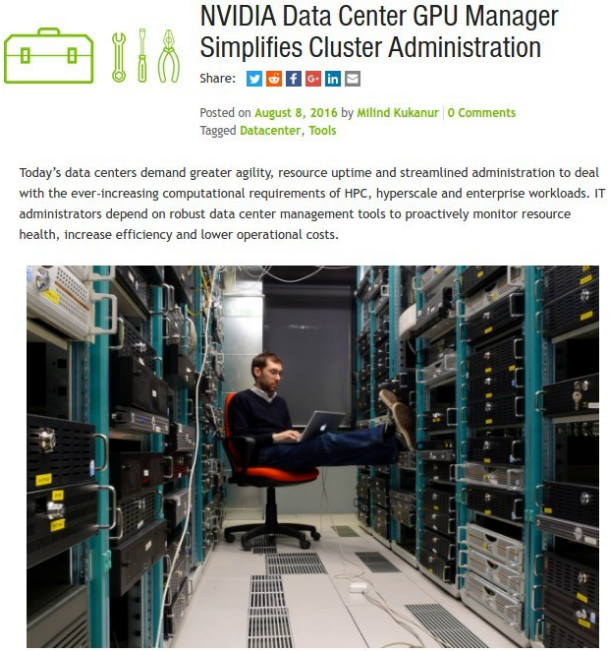
Local GPU Cluster?

- When individual workstation is not enough for your needs, you might want to think of building a cluster (if you also have the money for that)
- Buy multiple workstations each with 1-4 GPUs and connect
  - Gigabit Ethernet will be bottleneck
    - NVIDIA GPUDirect RDMA as solution



NVIDIA GPUDirect RDMA can bypass the CPU for inter-node communication – data is directly transfered between two GPUs.

timdettmers.com



NVIDIA Data Center GPU Manager Simplifies Cluster Administration

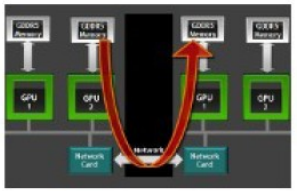
Share:

Posted on August 8, 2016 by Milind Kukanur | 0 Comments
Tagged Datacenter, Tools

Today's data centers demand greater agility, resource uptime and streamlined administration to deal with the ever-increasing computational requirements of HPC, hyperscale and enterprise workloads. IT administrators depend on robust data center management tools to proactively monitor resource health, increase efficiency and lower operational costs.

devblogs.nvidia.com

NVIDIA GPUDirect RDMA



Benchmarking GPUDirect RDMA on Modern Server Platforms

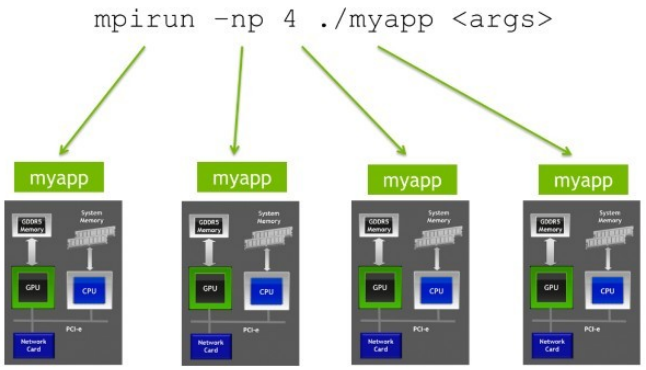
Share:

Posted on October 7, 2014 by Davide Rossetti | 9 Comments
Tagged Cluster, GPUDirect, MPI, Multi-GPU, RDMA

NVIDIA GPUDirect RDMA is a technology which enables a direct path for data exchange between the GPU and third-party peer devices using standard features of PCI Express. Examples of third-party devices include network interfaces, video acquisition devices, storage adapters, and medical equipment. Enabled on Tesla and Quadro-class GPUs, GPUDirect RDMA relies on the ability of NVIDIA GPUs to expose portions of device memory on a PCI Express Base Address Register region (BAR. See this white paper for more technical details).

Both Open MPI and MVAPICH2 now support GPUDirect RDMA, exposed via CUDA-aware MPI. Since January 2014 the Mellanox Infiniband software stack has supported GPUDirect RDMA on Mellanox ConnectX-3 and Connect-IB devices. See this post on the Mellanox blog for a nice introduction to the topic.

devblogs.nvidia.com



What is CUDA-aware MPI?

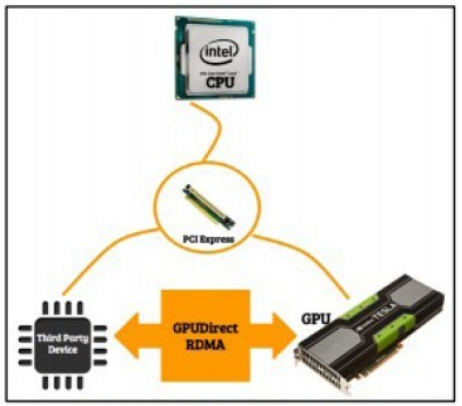
https://devblogs.nvidia.com/parallelforall/introduction-cuda-aware-mpi/
CUDA-aware MPI (the Message Passing Interface)

Mellanox MHZH29-XTR 40 Gb/s QSFP & 10GigE SFP+ Port ConnectX IB HCA Card

£98.63

Buy It Now
or Best Offer
+ £373.21 postage

http://www.mellanox.com/page/infiniband_cards_overview



How GPUDirect RDMA works?

www.univ-valenciennes.fr

GPUDirect RDMA over 40Gbps Ethernet

High Performance CUDA Clustering with Chelsio's T5 ASIC

Executive Summary

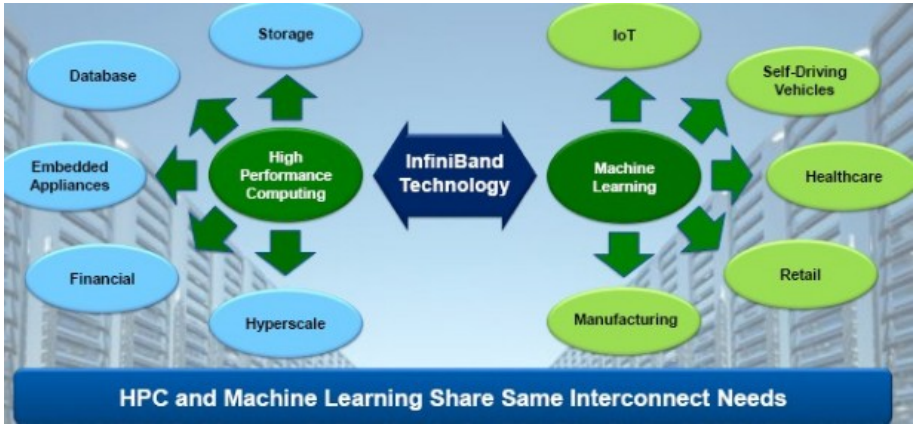
NVIDIA's GPUDirect technology enables direct access to a Graphics Processing Unit (GPU) over the PCI bus, shortcutting the host system and allows for high bandwidth, high message rate and low latency communication. When married to iWARP RDMA technology, high performance direct access to GPU processing units can be expanded seamlessly to Ethernet and Internet scales.

chelsio.com//T5-40Gb-Linux-GPUDirect.pdf

InfiniBand Fast networking **200 Gb/s**

THE NEXT PLATFORM

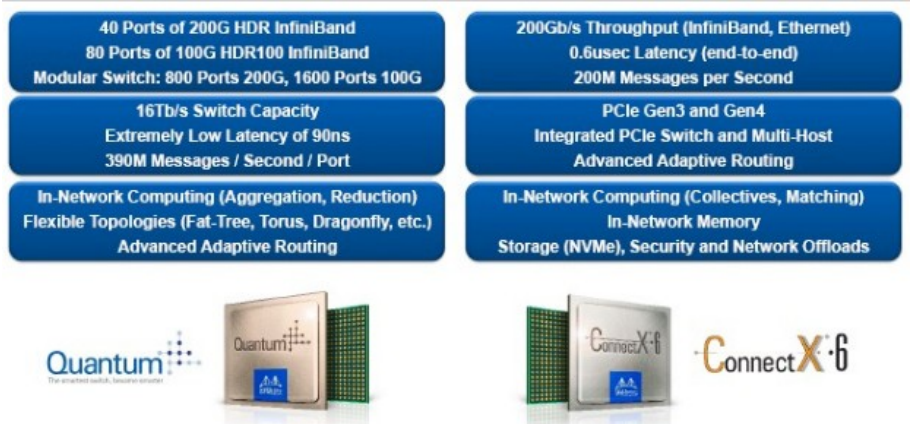
HOME COMPUTE STORE CONNECT CONTROL CODE

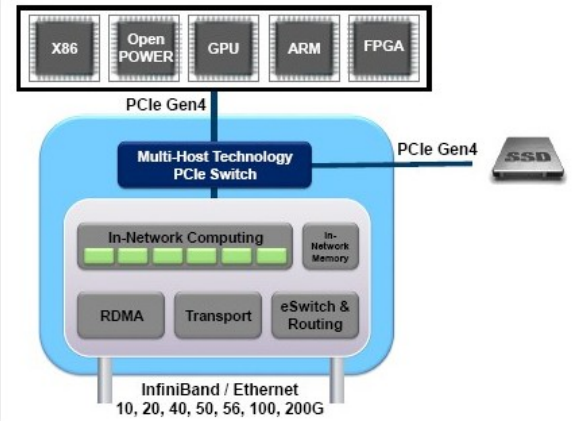


INFINIBAND BREAKS THROUGH THE 200G BARRIER

November 10, 2016 Timothy Prickett Morgan

https://www.nextplatform.com/2016/11/10/infiniband-breaks-200g-barrier/





"Baidu is using InfiniBand in its image recognition clusters, Facebook is using it in its Big Sur machine learning platform, PayPal is using it for its fraud detection systems, and even the DGX-1 machine learning appliance from Nvidia has four 100 Gb/sec adapters in it to keep those eight hungry "Pascal" Tesla P100 GPU accelerators well fed.

It is **hard for people to let anything but Ethernet into their networks**, but once they do, it probably gets a lot easier. The support of RDMA over Converged Ethernet (RoCE) mitigated this to a certain extent, but when HPC shops, hyperscalers, and cloud builders see they can get **200 Gb/sec InfiniBand** in 2017 and 400 Gb/sec InfiniBand in 2019, they might have a rethink."

Clustered Deep Learning SPARK #1

At the moment, hard to accelerate "single model" with cluster.

- Cluster good for ensembles / hyperparameter tuning → futureproof the setup so that it scales to Spark if needed?

Deep Learning with Apache Spark and TensorFlow

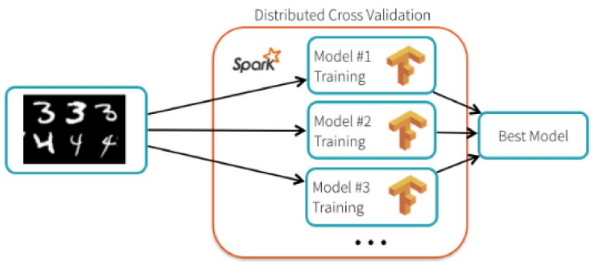
by Tim Hunter
Posted in **ENGINEERING BLOG** | January 25, 2016

HTTPS://DATABRICKS.COM/BLOG/2016/

You might be wondering: what's Apache Spark's use here when most high-performance deep learning implementations are single-node only? To answer this question, we walk through two use cases and explain how you can use Spark and a cluster of machines to improve deep learning pipelines with TensorFlow:

1. **Hyperparameter Tuning:** use Spark to find the best set of hyperparameters for neural network training, leading to 10X reduction in training time and 34% lower error rate.
2. **Deploying models at scale:** use Spark to apply a trained neural network model on a large amount of data.

The interesting thing here is that even though TensorFlow itself is not distributed, the hyperparameter tuning process is "embarrassingly parallel"



DeepSpark: Spark-Based Deep Learning Supporting Asynchronous Updates and Caffe Compatibility

Hanjoo Kim, Jaehong Park, Jaehee Jang, Sungroh Yoon

(Submitted on 26 Feb 2016 (v1), last revised 8 Mar 2016 (this version, v2))

HTTP://ARXIV.ORG/ABS/1602.08191

HTTP://DEEPSPARK.SNU.AC.KR/
HTTPS://GITHUB.COM/DEEPSPARK/DEEPSPARK

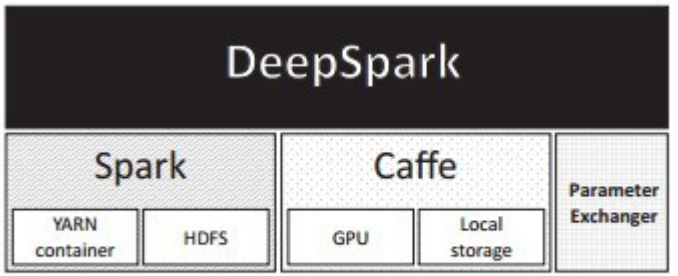


Figure 1: Stack diagram of DeepSpark architecture.

databricks

DATABRICKS.COM

The Unreasonable Effectiveness of Deep Learning on Apache Spark

by Miles Yucht and Reynold Xin
Posted in **ENGINEERING BLOG** | April 1, 2016

For the past three years, our smartest engineers at Databricks have been working on a stealth project. Today, we are unveiling **DeepSpark**, a major new milestone in Apache Spark. DeepSpark **uses cutting-edge neural networks to automate the many manual processes of software development**, including writing test cases, fixing bugs, implementing features according to specs, and reviewing pull requests (PRs) for their correctness, simplicity, and style.

Clustered Deep Learning SPARK #2

CaffeOnSpark Open Sourced for Distributed Deep Learning on Big Data Clusters

By Andy Feng(@afeng76), Jun Shi and Mridul Jain (@mridul_jain), Yahoo Big ML Team

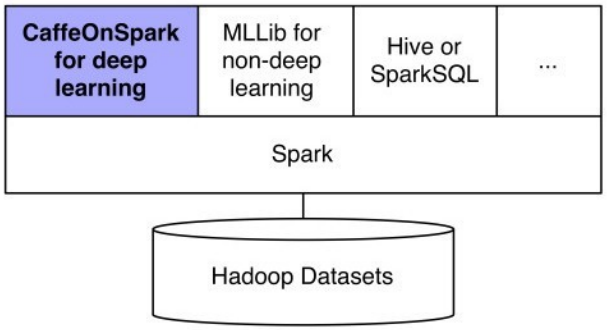


Figure 3: CaffeOnSpark as a Spark Deep Learning package

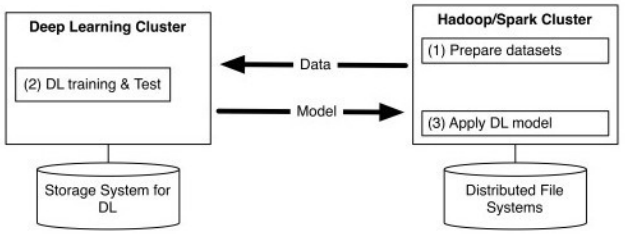


Figure 1: ML Pipeline with multiple programs on separated clusters

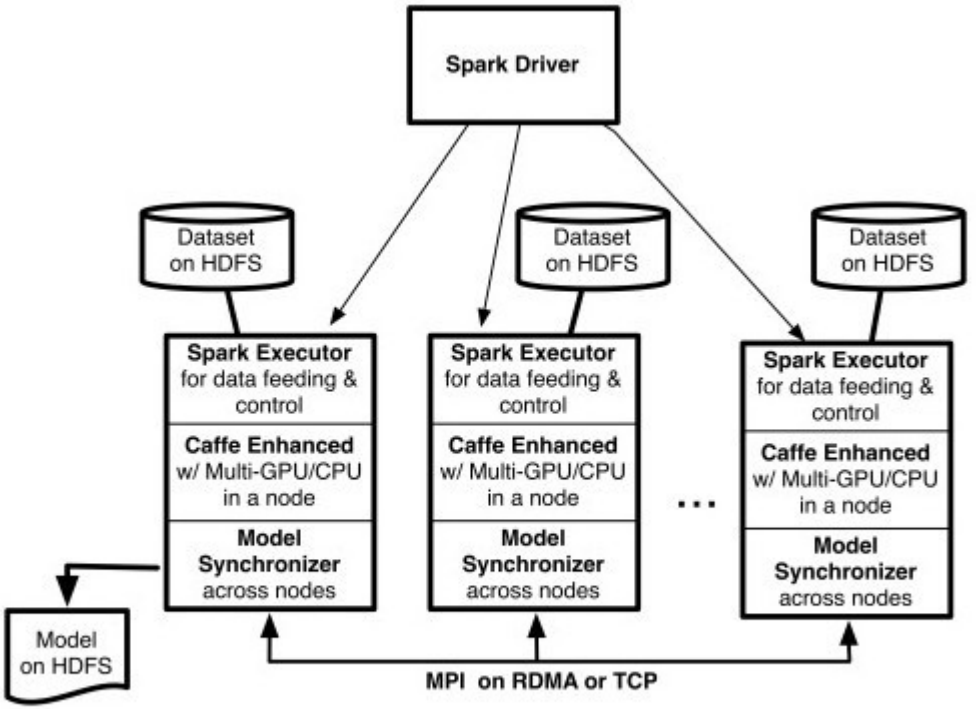
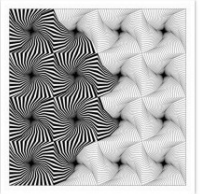


Figure 5: System Architecture
yahoonadoop.tumblr.com

CONVERGENCE COMING FOR SUPERCOMPUTING, MACHINE LEARNING

November 20, 2015 Nicole Hemsoth



When it comes to traditional supercomputing, the tools, frameworks, and software stacks tend to be codified, especially within the various domains that use high performance computing. In recent years, a new cadre of large-scale data analysis tooling has come into the HPC fold, but until more recently, machine learning, deep neural networks, and other re-emerging artificial intelligence tools have not been looped into the supercomputing world.

We have already described how machine learning and AI frameworks are using various elements of high performance computing, particularly on the accelerator side with companies like Nvidia dominating the computationally intensive training phase of machine learning via GPU acceleration and others, including FPGA maker, Xilinx, talking up the role of FPGAs for the inference portion of such workloads. However, outside of the use of these frameworks for hyperscale web companies, the connection between research-centric HPC and machine learning is still resolving.

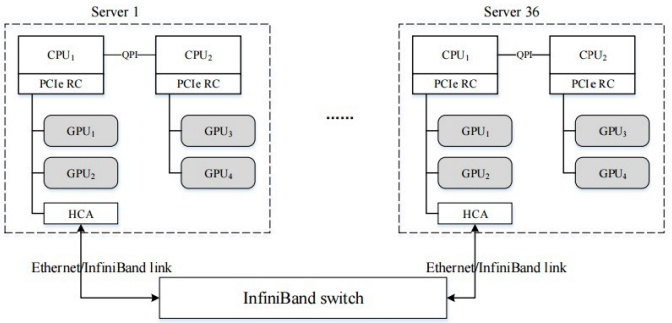
Another group of researchers tackled a more software-focused problem within machine learning at extreme scale in a presentation featuring a comparison of machine learning approaches for dealing with multi-collinearity in large-scale data analytics and data mining on high performance computing systems.

nextplatform.com/2015/11/20

Clustered Deep Learning OTHER OPTIONS

Deep Image: Scaling up Image Recognition

Ren Wu[1], Shengen Yan, Yi Shan, Qingqing Dang, Gang Sun
Baidu Research



Hardware Architecture. Each server has four Nvidia K40m GPUs and one InfiniBand adapter. Custom-built supercomputer, which we call Minwa is comprised of 36 server nodes, each with 2 six-core Intel **Xeon E5-2620** processors. Each sever contains **4 Nvidia Tesla K40m** GPUs and one FDR InfiniBand (56Gb/s) which is a high-performance low-latency interconnection and supports RDMA

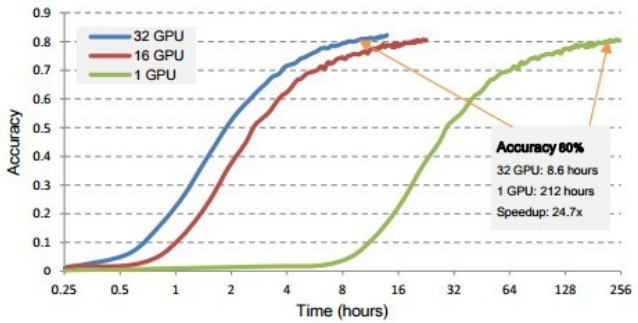


Figure 4: Validation set accuracy for different numbers of GPUs.

http://arxiv.org/abs/1501.02876

FireCaffe: near-linear acceleration of deep neural network training on compute clusters

Forrest N. Iandola, Khalid Ashraf, Matthew W. Moskewicz, Kurt Keutzer

(Submitted on 31 Oct 2015 (v1), last revised 8 Jan 2016 (this version, v2))

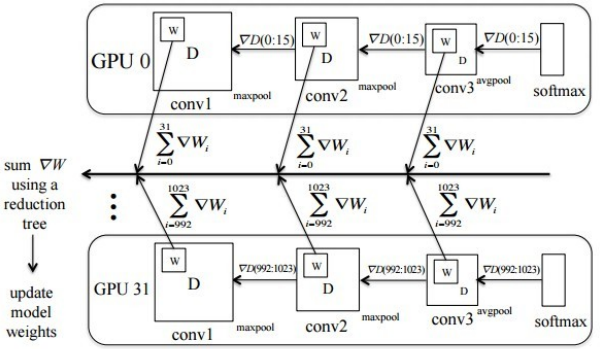


Figure 1. **Data parallel DNN training in FireCaffe:** Each worker (GPU) gets a subset of each batch.

Table 1. Volumes of data and computation for four widely-used DNN architectures. The batch size impacts all numbers in this table except for $|W|$, and we use a batch size of 1024 in this table. Here, TFLOPS is the quantity of computation to perform.

DNN architecture	typical use-case	data_size $\|D\|$	weight_size $\|W\|$	data/weight ratio	Forward+Backward TFLOPS/batch
NiN [32]	computer vision	5800MB	30MB	195	6.7TF
AlexNet [28]	computer vision	1680MB	249MB	10.2	7.0TF
GoogLeNet [41]	computer vision	19100MB	54MB	358	9.7TF
VGG-19 [39]	computer vision	42700MB	575MB	71.7	120TF
MSFT-Speech [38]	speech recognition	74MB	151MB	0.49	0.00015TF

Table 2. Accelerating the training of midsized deep models on ImageNet-1k.

	Hardware	Net	Epochs	Batch size	Initial Learning Rate	Train time	Speedup	Top-1 Accuracy
Caffe [27]	1 NVIDIA K20	AlexNet [29]	100	256	0.01	6.0 days	1x	58.9%
Caffe	1 NVIDIA K20	NiN [32]	47	256	0.01	5.8 days	1x	58.9%
Google cuda-convnet2 [28]	8 NVIDIA K20s (1 node)	AlexNet	100	varies	0.02	16 hours	7.7x	57.1%
FireCaffe (ours)	32 NVIDIA K20s (Titan supercomputer)	NiN	47	256	0.01	11 hours	13x	58.9%
FireCaffe-batch1024 (ours)	32 NVIDIA K20s (Titan supercomputer)	NiN	47	1024	0.04	6 hours	23x	58.6%
FireCaffe-batch1024 (ours)	128 NVIDIA K20s (Titan supercomputer)	NiN	47	1024	0.04	3.6 hours	**39x**	58.6%

Table 3. Accelerating the training of ultra-deep, computationally intensive models on ImageNet-1k.

	Hardware	Net	Epochs	Batch size	Initial Learning Rate	Train time	Speedup	Top-1 Accuracy	Top-5 Accuracy
Caffe	1 NVIDIA K20	GoogLeNet [41]	64	32	0.01	21 days	1x	68.3%	88.7%
FireCaffe (ours)	32 NVIDIA K20s (Titan supercomputer)	GoogLeNet	72	1024	0.08	23.4 hours	20x	68.3%	88.7%
FireCaffe (ours)	128 NVIDIA K20s (Titan supercomputer)	GoogLeNet	72	1024	0.08	10.5 hours	**47x**	68.3%	88.7%

http://arxiv.org/abs/1511.00175

Theano-MPI: a Theano-based Distributed Training Framework

He Ma[1], Fei Mao[2], and Graham W. Taylor[1]
[1] School of Engineering, University of Guelph, CA {hma02,gwtaylor}@uoguelph.ca
[2] SHARCNET, Compute Canada, CA feimao@sharcnet.ca

http://arxiv.org/abs/1605.08325

Poseidon: A System Architecture for Efficient GPU-based Deep Learning on Multiple Machines

Hao Zhang, Zhiting Hu, Jinliang Wei, Pengtao Xie, Gunhee Kim, Qirong Ho, Eric Xing

(Submitted on 19 Dec 2015)

http://arxiv.org/abs/1512.06216

Strategies and Principles of Distributed Machine Learning on Big Data

Eric P. Xing, Qirong Ho, Pengtao Xie, Wei Dai

(Submitted on 31 Dec 2015)

http://arxiv.org/abs/1512.09295

Baidu 百度

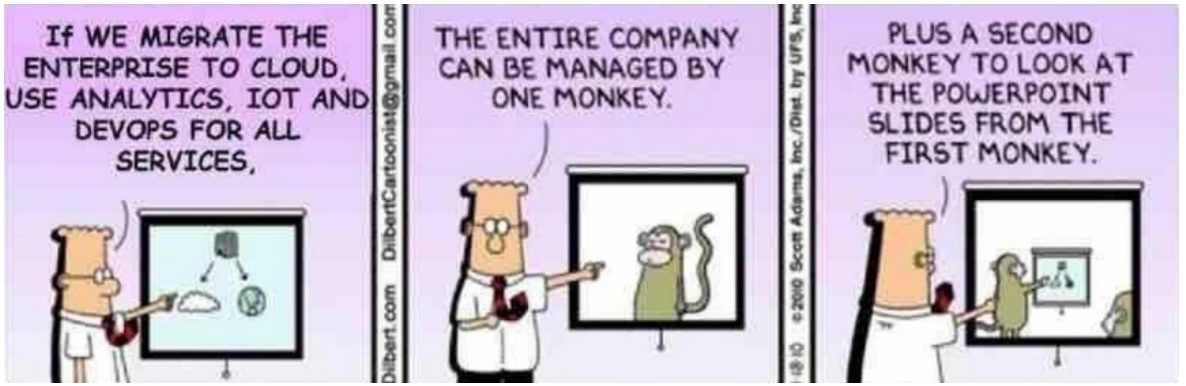
baidu / Paddle

Code | Issues 21 | Pull requests 4 | Wiki | Pulse | Graphs

PArallel Distributed Deep LEarning http://www.paddlepaddle.org/

https://github.com/baidu/paddle

Cloud Services



http://dilbert.com/search_results?terms=Cloud+Computing

Cloud GPU Training NVIDIA / AMAZON

- Not the most economical choice, but can be used to supplement the local workstation occasionally when needed.

GPU Cloud Computing

The following organizations offer cloud computing services to enable GPU computing from anywhere around the world.

Learn more about GPUs in the cloud, check out the GPU Technology Conference sessions on cloud computing.

http://www.nvidia.com/object/gpu-cloud-computing-services.html

Partner	Services Offered	Regions Supported
Amazon Web Services	• Hosted GPUs • GPU Cloud	• North America
Aliyun	• GPU Cloud • High Performance Computing	• Asia
Microsoft Azure	• Hosted GPUs • GPU Cloud	• North America
Outscale	• High Performance Computing • GPUs on Demand	• North America • Europe • Asia
Peer 1 Hosting	• Hosted GPUs • GPU Cloud	• North America • Europe
Penguin Computing	• Hosted GPUs	• North America
RapidSwitch	• High Performance Computing • GPU cloud	• Europe
Rescale	• High Performance Computing • GPUs on Demand	• North America • Europe • Asia
SoftLayer	• High Performance Computing • Hosted GPUs • GPUs on Demand	• North America • Europe • Asia

Amazon EC2 Instance Types

GPU

G2

This family includes G2 instances intended for graphics and general purpose GPU compute applications.

Features:

- High Frequency Intel Xeon E5-2670 (Sandy Bridge) Processors
- High-performance NVIDIA GPUs, each with 1,536 CUDA cores and 4GB of video memory

Model	GPUs	vCPU	Mem (GiB)	SSD Storage (GB)
g2.2xlarge	1	8	15	1 x 60
g2.8xlarge	4	32	60	2 x 120

Use Cases

On-demand
LINUX

GPU Instances - Current Generation					
g2.2xlarge	8	26	15	60 SSD	$0.702 per Hour
g2.8xlarge	32	104	60	2 x 120 SSD	$2.808 per Hour

Spot
LINUX

GPU Instances - Current Generation	
g2.2xlarge	$0.1525 per Hour
g2.8xlarge	$0.9125 per Hour

https://aws.amazon.com/ec2/pricing/

Cloud GPU Training AMAZON

- Amazon has become *de facto* standard or a synonym for cloud computing while in reality their GPU instances are quite under-powered and not the best choice for deep learning:
  - Amazon: *"you'll benefit from using GPU instances, which provide access to NVIDIA GPUs with up to 1,536 CUDA cores and 4 GB of video memory."*

    G2 instances use NVIDIA Grid K520 GPUs with one GPU giving ~2,45 TFLOPS (FP32)"
- To put this into context, our chosen GPU Pascal-based Titan X (£1,000/$1,200) gives around **11 TFLOPS** (FP32) with 12 GB compared to the 16 GB of 4 GPUs in g2.8 instance (~**10 TFLOPS**).
  - The on-demand hourly price for g2.8 instance is $2.808, amortizing the Titan X investment in ~388 hours (10TFLOPS/11TFLOPS * $1,200 / $2.808/h) or in **16.2 days (!)** with constant use (excluding electricity prices).
    - With electricity: $0 = c(tEP+p) - tA \rightarrow$ $t = -cp / (cEP-A)$ ~395 hours

      Where *t* is the amortization time (hours), *c* equates FLOPS per hour, *E* price of electricity (£0.098→$0.13), *P* computer power usage (0.4 kW with single Titan X + CPU), *A* price of Amazon instance per hour ($2.808/hour). This gives a **ballpark estimate**.
  - Monthly Amazon cost of **constant use** would be a staggering **$2,021** without using the cheaper and more unreliable spot instance.
- With its current state, the Amazon Cloud **does not seem to be really appealing** either in performance or cost-sense.

Cloud GPU Training AMAZON UPDATE SEPT 2016

Amazon finally addresses their underpowered Cloud services and respond to the pressure by Azure which hopefully leads to more customer-friendly pricing as well

AWS Blog

New P2 Instance Type for Amazon EC2 – Up to 16 GPUs

by Jeff Barr (@jeffbarr) | on 29 SEP 2016 | in Amazon EC2, Launch

https://aws.amazon.com/blogs/aws/new-p2-instance-type-for-amazon-ec2-up-to-16-gpus/

Instance Name	GPU Count	vCPU Count	Memory	Parallel Processing Cores	GPU Memory	Network Performance
p2.large	1	4	61 GiB	2,496	12 GB	High
p2.8xlarge	8	32	488 GiB	19,968	96 GB	10 Gigabit
p2.16xlarge	16	64	732 GiB	39,936	192 GB	20 Gigabit

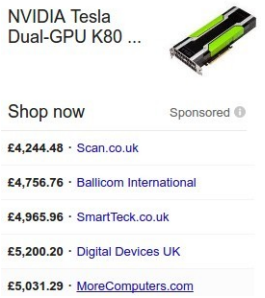
This new instance type incorporates up to 8 NVIDIA Tesla K80 Accelerators, each running a pair of NVIDIA GK210 GPUs. Each GPU provides 12 GB of memory (accessible via 240 GB/second of memory bandwidth), and 2,496 parallel processing cores.

All of the instances are powered by an AWS-Specific version of Intel's Broadwell processor, running at 2.7 GHz. The p2.16xlarge gives you control over C-states and P-states, and can turbo boost up to 3.0 GHz when running on 1 or 2 cores.

The GPUs support CUDA 7.5 and above, OpenCL 1.2, and the GPU Compute APIs. The GPUs on the p2.8xlarge and the p2.16xlarge are connected via a common PCI fabric. This allows for low-latency, peer to peer GPU to GPU transfers.

TESLA K80 ACCELERATOR FEATURES AND BENEFITS

- Launched back in 2014
- Up to 2.91 Teraflops double-precision performance with NVIDIA GPU Boost
- Up to 8.73 Teraflops single-precision performance with NVIDIA GPU Boost
  http://www.nvidia.com/object/tesla-k80.html#sthash.BweuABKL.dpuf



Tesla Model	K10	K20	K20X	K40	K80	M4	M40	P100
GPU	2 * GK104	GK110	GK110	GK110B	2 * GK210B	GM206	GM200	GP100
CUDA Cores	2 * 1,536	2,496	2,688	2,880	4,992	1,024	3,072	3,584
Base Core Clock Speed	745 MHz	706 MHz	732 MHz	745 MHz	560 MHz	872 MHz	948 MHz	1,328 MHz
GPU Boost Clock Speed	–	–	–	875 MHz	875 MHz	1,072 MHz	1,114 MHz	1,480 MHz
SMXs or SMMs	2 * 8	13	14	15	2 * 13	8	24	56
Base HP, Teraflops	–	–	–	–	–	–	–	*
Peak HP, Teraflops	–	–	–	–	–	–	–	21.2
Base SP, Teraflops	4.58	3.52	3.95	4.29	5.6	*	*	*
Peak SP, Teraflops	4.58	3.52	3.95	5.0	8.74	2.2	7.0	10.6
Base DP, Teraflops	0.19	1.17	1.31	1.43	1.87	*	*	*
Peak DP, Teraflops	0.19	1.17	1.31	1.66	2.91	0.06	0.20	5.30
GDDR5/HBM2 Memory	8 GB	5 GB	6 GB	12 GB	24 GB	4 GB	24 GB	16 GB
Memory Clock Speed	2.5 GHz	2.6 GHz	2.6 GHz	3.0 GHz	2.5 GHz	2.75 GHz	3.0 GHz	-
Memory Bandwidth	320 GB/sec	208 GB/sec	250 GB/sec	288 GB/sec	480 GB/sec	88 GB/sec	288 GB/sec	720 GB/sec
Power Draw	225 W	225 W	235 W	235 W	300 W	50 W – 75 W	250 W	300 W
HP Efficiency (Gigaflops/Watt)	–	–	–	–	–	–	–	70.7
SP Efficiency (Gigaflops/Watt)	20.4	15.6	16.8	21.3	29.1	29.3	28.0	35.3
DP Efficiency (Gigaflops/Watt)	0.8	5.2	5.6	7.1	9.7	0.8	0.8	17.7

* *Base HP, SP, and DP teraflops unknown*

Nvidia Not Sunsetting Tesla Kepler And Maxwell GPUs Just Yet
http://www.nextplatform.com/2016/04/07

Cloud GPU Training MICROSOFT AZURE

- Microsoft Azure has recently launched beta GPU cloud instances with competitive specs (well if you are into FP64 training):

BLOG > ANNOUNCEMENTS , VIRTUAL MACHINES

Azure N-Series preview availability

Posted on August 4, 2016

Corey Sanders, Director of Program Management, Azure

Following are the Tesla K80 GPU sizes available (2.91 Teraflops of double-precision and up to 8.93 Teraflops of single-precision performance.):

	NC6	NC12	NC24
Cores	6 (E5-2690v3)	12 (E5-2690v3)	24 (E5-2690v3)
GPU	1 x K80 GPU (1/2 Physical Card)	2 x K80 GPU (1 Physical Card)	4 x K80 GPU (2 Physical Cards)
Memory	56 GB	112 GB	224 GB
Disk	380 GB SSD	680 GB SSD	1.44 TB SSD

Following are the Tesla M60 GPU Sizes available:

	NV6	NV12	NV24
Cores	6 (E5-2690v3)	12 (E5-2690v3)	24 (E5-2690v3)
GPU	1 x M60 GPU (1/2 Physical Card)	2 x M60 GPU (1 Physical Card)	4 x M60 GPU (2 Physical Cards)
Memory	56 GB	112 GB	224 GB
Disk	380 GB SSD	680 GB SSD	1.44 TB SSD

N-Series

The N-series is a family of Azure Virtual Machines with GPU capabilities. GPUs are ideal for compute and graphics-intensive workloads, helping customers to fuel innovation through scenarios such as high-end remote visualisation, deep learning and predictive analytics. Available in preview today, the N-series will feature the NVIDIA Tesla accelerated platform as well as NVIDIA GRID 2.0 technology, providing the highest-end graphics support available in the cloud today. Also, there is a second low latency, high-throughput network interface (RDMA) optimised VM configuration (NC24r) which is tuned for tightly coupled parallel computing workloads.

N-Series is not available in the Central US region. Please select another region.

N-series virtual machines are currently only available in the South Central US region.

Instance	Cores	Memory	Storage	GPU	Price
NV6	6	56 GB	340 GB	1 – M60	$0.73 per hour
NV12	12	112 GB	680 GB	2 – M60	$1.46 per hour
NV24	24	224 GB	1,440 GB	4 – M60	$2.92 per hour
NC6	6	56 GB	340 GB	1 – K80	$0.66 per hour
NC12	12	112 GB	680 GB	2 – K80	$1.33 per hour
NC24	24	224 GB	1,440 GB	4 – K80	$2.66 per hour
NC24r	24	224 GB	1,440 GB	4 – K80	$2.99 per hour

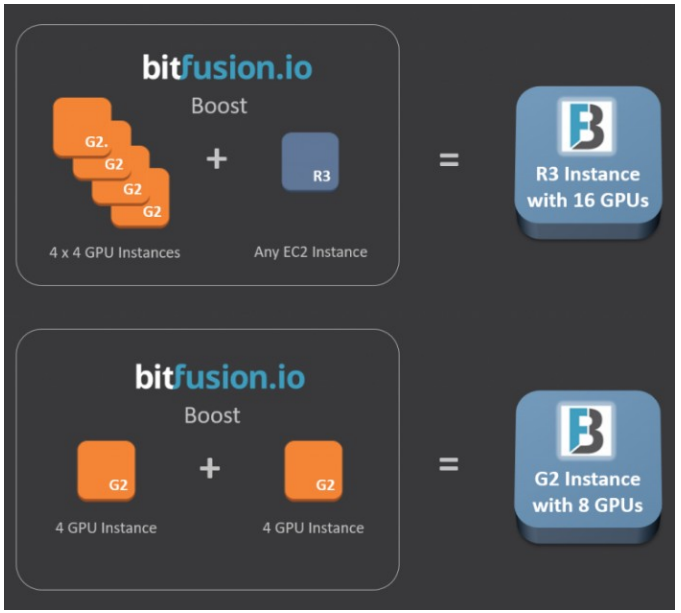
https://news.ycombinator.com/item?id=12267961

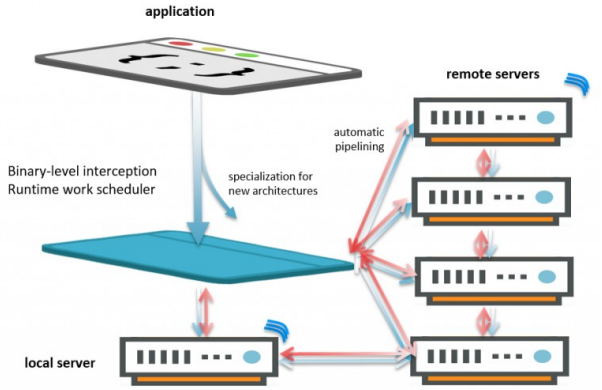
Cloud GPU Training NIMBIX WITH AMAZON

Introducing Monster Machines, the world's largest cloud GPU instances on AWS

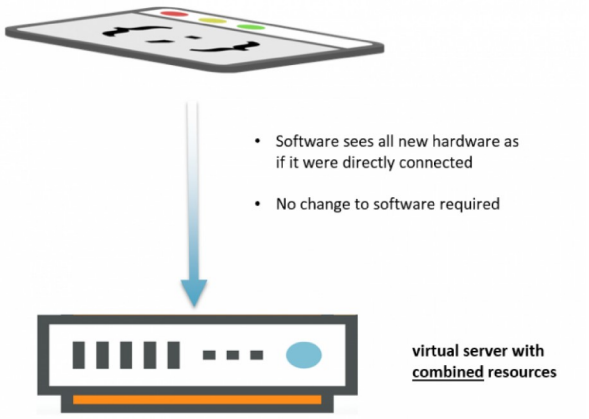
The largest Amazon EC2 instance, the g2.8xlarge, currently maxes out at just 4 GPUs. With a combination of our Bitfusion Boost remoting technology and the use of CloudFormation templates (step-by step Caffe CFN tutorial), we now allow you to easily spin up virtual instances with a lot more GPU power. For example, you can combine a large memory machine with an unlimited number of GPU nodes into a single more powerful virtual instance:

http://www.bitfusion.io/2016/03/31/introducing-monster-machines-worlds-largest-cloud-gpu-instances-aws/





The Boost runtime intercepts API calls, splits up the compute and data, and forwards the requests to a fast run-time scheduler to dispatch computation to both local and remote GPUs. As a result, you can combine the compute resources of various nodes into a **single giant node**. In fact, the GPU application doesn't see any of this complexity; it just sees itself running on a single giant machine as shown in the figure below.



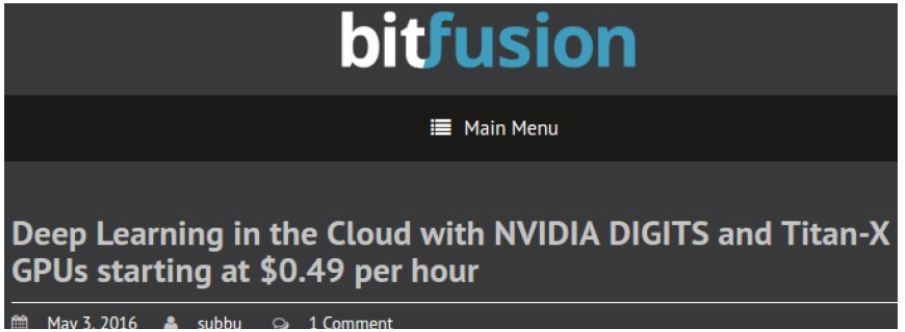
Feeling adventurous? Try out the massive R3 instance with 16 GPUs, 32 CPUs, and 244 GB of Memory:

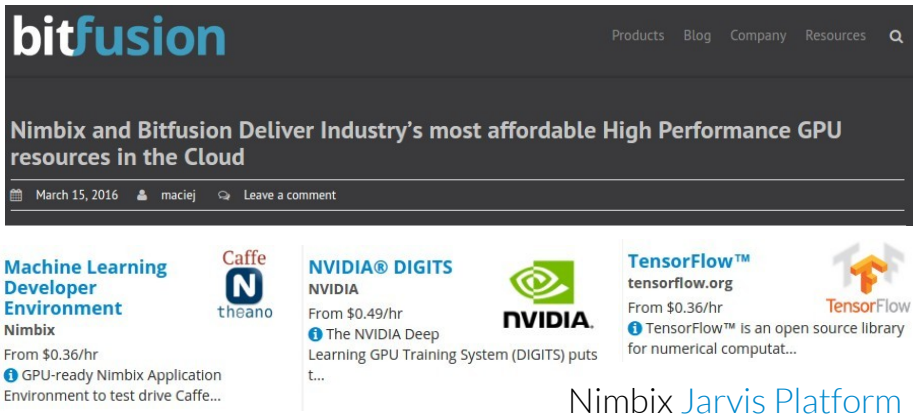
Launch R3 with 16 GPUs Instance

For deep learning, it would be nice to have P100-based GPU cloud with both FP16 and FP32 training instead of the Teslas. Or just use more extensively Titan X in cloud as well.

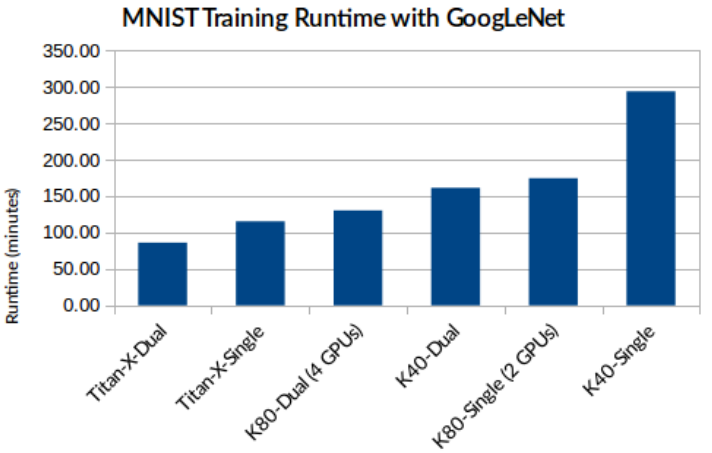
Eventually use Boost with **"monster" Titan X machines**?

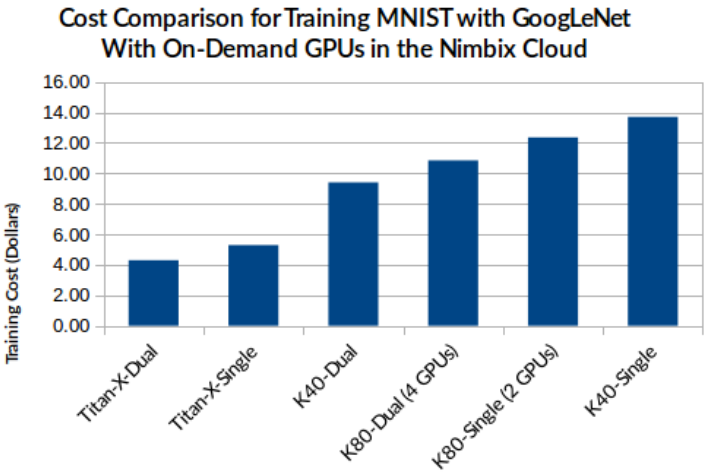
Cloud GPU Training NIMBIX #1





Nimbix Jarvis Platform





For GoogLeNet, dual GPU configurations have superior runtime to the same single GPU configuration of the same hardware in all cases. However, the Maxwell-architecture **Titan X still wins with the fastest runtime** at 85.95 minutes to train the MNIST data set. The best case training cost is also the best total runtime for GoogLeNet, which is the **dual Titan X** configuration at $4.30.

www.nimbix.net

github.com

Cloud GPU Training IBM

The partnership between Nvidia and IBM is giving Big Blue a leg up in terms of making a wider array of GPUs available to suit different workloads. Currently, IBM's cloud boasts the K80, as well as the lower power and less beefy K10. Today that suite of GPU options was enriched with the addition of the virtualization-ready Nvidia M60 cards, which can support a wider range of workloads—from HPC applications, to machine learning workloads, to virtual services and gaming platforms.

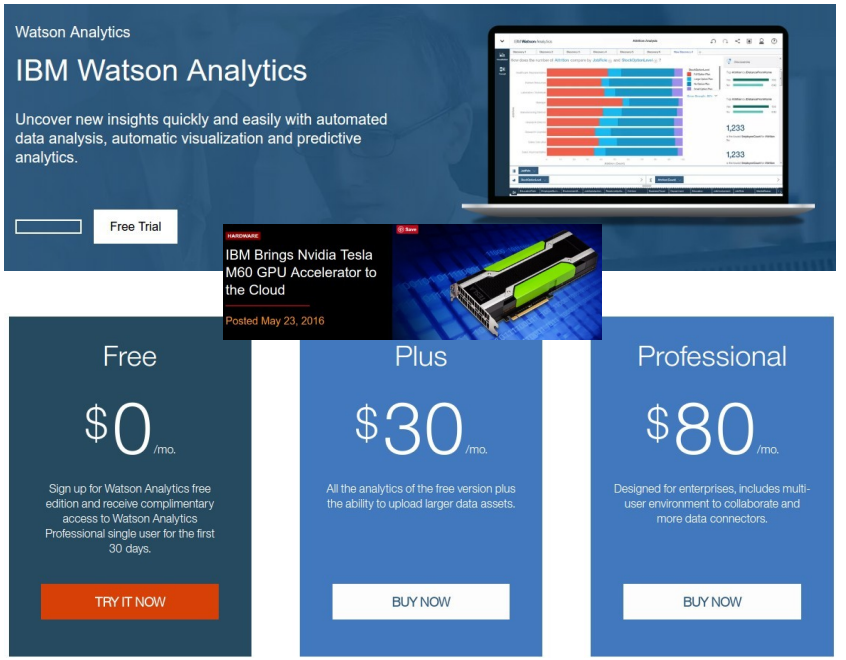
IBM is expecting that the M60 GPU will be useful for customers who want to move machine learning and deep learning training into the cloud, but he says for now, a lot of them have started by using the far cheaper Titan X GPUs ($1000 versus several thousand) as Baidu does. He says that while he knows many shops are using these consumer cards, the **Titan X can't be licensed for use in the cloud** (Nvidia's Marc Hamilton talks about that in more detail here), but the K80 is already serving production machine learning use cases and the M60 will do so as well.

http://www.nextplatform.com/2016/05/19/ibm-extends-gpu-cloud-capabilities-targets-machine-learning/

IBM Watson Analytics vs. Microsoft Azure Machine Learning (Part 1)

http://www.kdnuggets.com/2014/12/

www.kdnuggets.com/2015/04



https://watson.analytics.ibmcloud.com/pricing

Why no Titan X Cloud Services?

- The most cost-effective way to train deep neural nets in the cloud would be to use Titan X boards with their better FP32 performance compared to the Tesla range but they are still not really used. Why?

> *"Baidu's research team opted for the slightly less higher performance, but presumably **far cheaper Titan X GPUs**, which pack a whopping 6-plus teraflops peak into a $1,000 GPU package. .. Of course, **peak teraflops** versus **real-world performance** are different things entirely. Bryan Catanzaro says they are getting 3 teraflops out of their Titan X GPUs on their neural network code. And what is most interesting here, they have made yet another GPU computing leap with those, scaling first from four to eight GPUs in a single box, then going on to use sixteen GPUs per training run—an uncommon feat that Catanzaro says will gain traction elsewhere in time, especially once optimization tricks of the trade are passed down the line at other companies" - nextplatform.com/2015/12/11*

> *"At the time, Bryan Catanzaro told us about their use of Nvidia **Titan X GPU cards as the most cost efficient option** for the computationally-intensive task of model training, despite the availability of other GPUs, including the M40 and for the inference phase, M4 as well as other more powerful GPUs, including the supercomputing oriented Tesla K80." - nextplatform.com/2016/04/22*

- It is possible to deploy Titan X's as bare metal cloud, but not as virtualised cloud due to NVIDIA licensing policies. It is possible to go around this in your own virtualized solution e.g. using this "non-root GPU passthrough setup" by Ruslan Habalov

https://news.ycombinator.com/item?id=12267961 :

> dharma1 12 August 2016
> Is there any reason cloud providers don't offer virtualised Titan X, or GTX 1080?
>
> > dogma1138 12 August 2016
> >
> > As for the passthrough it is possible but it requires a few hacks and is technically in violations of NVIDIA's policies, and unlike the Tesla parts there is no full virtualization once the a device has been initialized it can only be assigned to 1 host/guest and a full host reboot is required to reinitialize the device again. You also need to do a few UEFI hacks to prevent the UEFI from initializing the GPU before your host OS loads and it can be passed through to the guest.
> >
> > This isn't something that cloud providers would dealing with, NVIDIA will not sell them anything but Tesla/Grid parts and buying from AIB's/OEM's and hacking your way through it isn't an option, you'll get zero support from NVIDIA, you could not thin provision your GPU's and you'll need to do a full host reboot every time you want to reassign the GPU or try to figure out if you can write a custom BIOS for your card that would somehow allow you to reinitialize it.

Cloud GPU Training RunAbove

- GTX 1080&1070 GPU Bare Metal Servers

Choose your server		
	Project 1521-6	**Project 1521-7**
CPU	Xeon E5-2630v3	
RAM	128GB DDR3	
Storage	1x240GB SSD	
GPU	1xGeForce GTX 1080	1xGeForce GTX 1070
Price	€69.99 excl. VAT/2 weeks (and then €199.00 excl. VAT/mo.)	€59.99 excl. VAT/2 weeks (and then €159.99 excl. VAT/mo.)

https://www.runabove.com/titan-x-gpu-servers.xml

GIGABYTE (71)

GIGABYTE GeForce GTX 1080 G1 Gaming GV-N1080G1 GAMING-8GD Video Card

Free in-game value of Paragon w/ purchase, limited offer

$649.99 (2 Offers)

GIGABYTE (16)

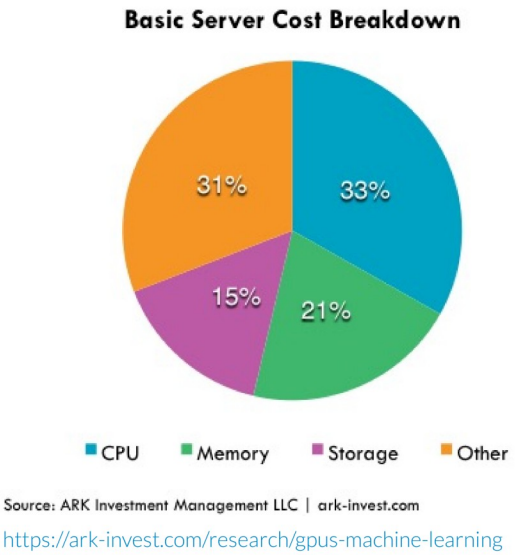
GIGABYTE GeForce GTX 1070 Founders Edition GV-N1070D5-8GD-B

Free in-game value of Paragon w/ purchase, limited offer

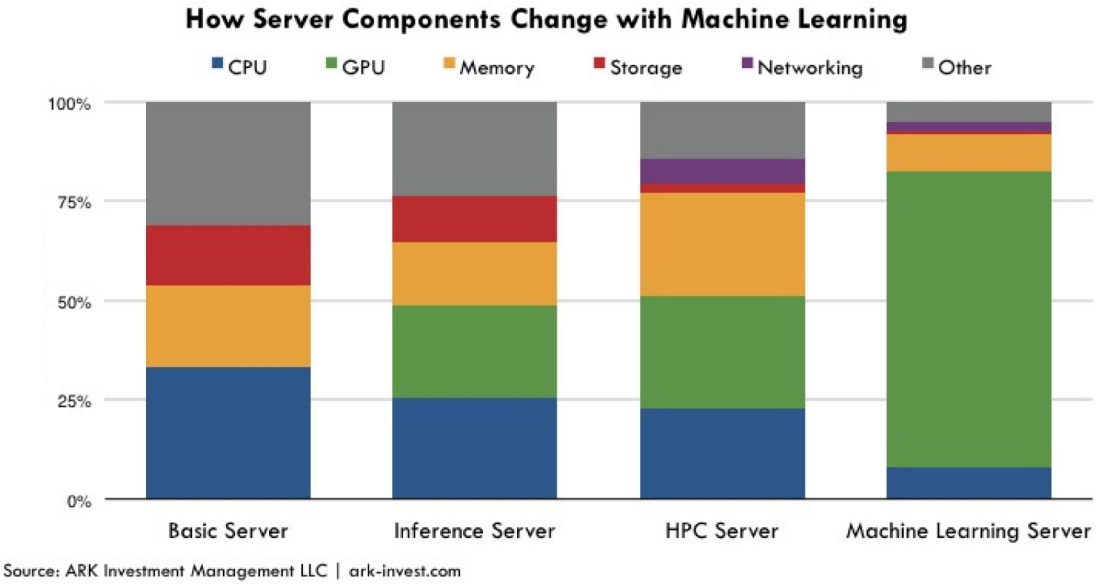
$449.99

Cloud GPU Services | Inference

- The cost structure is quite different if you can train the network locally in your office, and only need cloud-solutions for the deployment (e.g. Apple Siri)



https://ark-invest.com/research/gpus-machine-learning



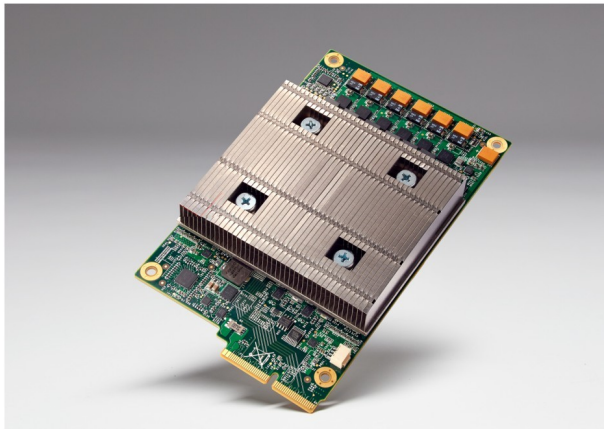
- If you are developing a product where your target customer do not have the computing powers to do the inference, but the internet bandwidth, the cloud-based inference is an attractive alternative for local NVIDIA GPUs.

Cloud GPU Services | Inference #2

- Now with inference, there are other options to NVIDIA (although NVIDIA stays as the easiest solution)

Google is bringing custom tensor processing units to its public cloud

JORDAN NOVET MAY 18, 2016 12:10 PM
TAGS: ALPHABET, GOOGLE, GOOGLE I/O 2016



venturebeat.com
forbes.com

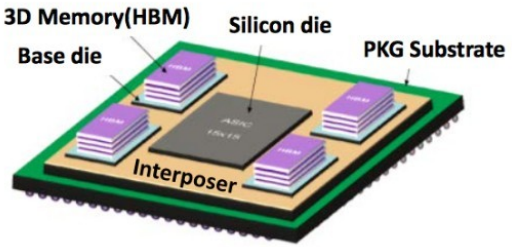
Google Cloud Platform

Need Deep Learning? There's a Cloud for That.

by Stacey Higginbotham @gigastacey FEBRUARY 29, 2016, 9:00 AM EDT

Today Nervana's cloud is based on graphics processors purchased from Nvidia, but the founders of Nervana hope to replace the underlying hardware by the end of 2016 with specialized chips of their own design. Until

fortune.com



The Nervana Engine (coming in 2017) is an application specific integrated circuit (ASIC) that is custom-designed and optimized for deep learning, resulting in a 10x increase in training speed and ensuring that Nervana Cloud will remain the world's fastest deep learning platform for the foreseeable future!
www.nervanasys.com

GoogleNet V1 - Input 128x3x224x224

Library	Class	Time (ms)	forward (ms)	backward (ms)
Nervana-neon-fp16	ConvLayer	230	72	157
Nervana-neon-fp32	ConvLayer	270	84	186
TensorFlow	conv2d	445	135	310
CuDNN[R4]-fp16 (Torch)	cudnn.SpatialConvolution	462	112	349

https://github.com/soumith/convnet-benchmarks

IEEE SPECTRUM

Now You Too Can Buy Cloud-Based Deep Learning

Cloud-computing services deliver AI to the rest of us

By Jeremy Hsu
Posted 27 Jul 2016 | 15:00 GMT

The Ivy League of Deep Learning

	Cloud machine-learning service	Open-source machine-learning tools	Deep-learning startup acquisitions
Amazon	Amazon Machine Learning	**DSSTNE** Deep Scalable Sparse Tensor Network Engine; library for building deep-learning models	Orbeus
Facebook	None	Tools for deep-learning models released through the open-source Torch library	Wit. AI
Google	Google Cloud Machine Learning	**TensorFlow** Library for developing deep-learning models and more general machine-learning models	Dark Blue Labs, DeepMind, DNNresearch, Moodstocks, Vision Factory
IBM	IBM Watson Analytics	**IBM SystemML** Optimization platform for general machine-learning models on the open-source Apache Spark library	AlchemyAPI
Microsoft	Microsoft Azure Machine Learning	**CNTK** Computational Network Toolkit; library for building deep-learning models	SwiftKey

spectrum.ieee.org

Google Cloud GPU Updated Nov 16

Announcing GPUs for Google Cloud Platform Tuesday, November 15, 2016. Early in 2017, Google Cloud Platform will offer GPUs worldwide for Google Compute Engine and Google Cloud Machine Learning users. Google Cloud will offer AMD FirePro S9300 x2 that supports powerful, GPU-based remote workstations. We'll also offer NVIDIA® Tesla® **P100** and K80 GPUs for deep learning, AI and HPC applications that require powerful computation and analysis.

GPUs on Google Cloud

AMD FirePro S9300 x2

Finally a cloud with 'deep learning GPU' with both FP16 and FP32 computations available

NVIDIA Tesla P100

5.3 TFLOPS double (FP64)
10.6 TFLOPS single (FP32)
21.2 TFLOPS half (FP16)

NVIDIA Tesla K80s

2.9 TFLOPS double (FP64)
4.29 TFLOPS single (FP32)

NAS SERVER

Data Storage Solution

LOCAL RAID (NOT-NAS) OR NAS SERVER

SUMMARY

Local desktop RAID selected above EXTENSIBLE NOW TO NAS IF YOU WANT

- 24 TB (4 x 6 TB, 7200rpm) with RAID 5 gives 18 TB of storage
  £960 at Amazon.co.uk
- 32 TB (4 x 8 TB, 7200rpm) with RAID 5 gives 24 TB of storage
  £1,720 at scan.co.uk

Commercial small-business NAS Server (if needed)

- Synology DiskStation DS1515+ for 5 HDDs (£550 extra)

Additional cloud support

- Dropbox-like use (alphr.com)
- Cloud backup of the backup $4-8/month (http://cloudwards.net/)

RAID Levels

ComputerWeekly.com | E-guide

Storage trends 2016: Storage priorities, cloud appliances and NAS vs object storage

SSD RAID essentials: What you need to know about flash and Raid

With the advent of flash and high-capacity HDDs, what do today's storage professionals need to know about Raid and its new variants?

RAID Levels Explained

BY SAMARA LYNN 27 MAR 2014, 5:01 A.M.

If you've ever looked into purchasing a **NAS device** or **server**, particularly for a small business, you've no doubt come across the term "RAID." RAID stands for Redundant Array of Inexpensive (or sometimes "Independent") Disks. In general, a RAID-enabled system uses two or more hard disks to improve the performance or provide some level of fault tolerance for a machine—typically a NAS or server. Fault tolerance simply means providing a safety net for failed hardware by ensuring that the machine with the failed component, usually a hard drive, can still operate. Fault tolerance lessens interruptions in productivity, and it also decreases the chance of data loss.

RAID Level Comparison

Features	RAID 0	RAID 1	RAID 1E	RAID 5	RAID 5EE	RAID 6	RAID 10
Minimum # Drives	2	2	3	3	4	4	4
Data Protection	No Protection	Single-drive failure	Single-drive failure	Single-drive failure	Single-drive failure	Two-drive failure	Up to one disk failure in each sub-array
Read Performance	High	High	High	High	High	High	High
Write Performance	High	Medium	Medium	Low	Low	Low	Medium
Read Performance (degraded)	N/A	Medium	High	Low	Low	Low	High
Write Performance (degraded)	N/A	High	High	Low	Low	Low	High
Capacity Utilization	100%	50%	50%	67% - 94%	50% - 88%	50% - 88%	50%
Typical Applications	High end workstations, data logging, real-time rendering, very transitory data	Operating system, transaction databases	Operating system, transaction databases	Data warehousing, web serving, archiving	Data warehousing, web serving, archiving	Data archive, backup to disk, high availability solutions, servers with large capacity requirements	Fast databases, application servers

Upgrade to NAS server

- Now we have assumed that you only need one workstation and you would not really benefit that much from NAS server.
- However if your startup team / research lab starts to grow, and more people need to access the same data it might make sense to get a NAS server for the data
- Now in this case, you can simply re-use the already bought “NAS HDD”, and just get the NAS server.
  - So in practice there has not been any redundant investments and the choice of starting with a local RAID seems good as it is easily upgradable to a NAS server as the most expensive component in that scenario tends to be the hard drives itself.

LOCAL RAID → NAS SERVER

40 TB (32 TB with RAID 5, and with Hybrid Raid you have more flexibility)

Synology DiskStation DS1515+

●●●●○ EDITOR RATING: EXCELLENT (4.0)

REVIEW | COMMENTS | SPECS

*Capable of holding up to **40TB of data** (with five 8TB drives) or scaling up to 120TB (with two of Synology's optional **DX513 expansion units** plugged into the eSATA ports on the back panel)*

PROS
Easy administration, particularly of AES-256-encrypted folders and private cloud access. Four Ethernet ports. Quiet operation.

CONS
PC backup can be complicated. No status-monitor LCD or push-button USB drive backup.

BOTTOM LINE
The five-bay Synology DiskStation DS1515+ network-attached storage (NAS) device will impress SMBs seeking a secure solution for data and app serving and private cloud creation.

uk.pcmag.com/synology-diskstation-ds1515

1x
£550

+ 5x
£2,150 40 TB
£1,200 30 TB

6 TB 7200rpm, ~£240
5400rpm, ~£207

8 TB 7200rpm, ~£430
5400rpm, ~£300

www.wdc.com

= £1,750 30 TB - £40.00/TB
= £2,700 40 TB - £53.75/TB

Can be extended to a total of 120 TB

2 x Expansion units (DX513)

For 120 TB, total cost = **£7,700**
= £2,700 + 2*£2,500

40 TB

	Product Name	Unit Price
	Synology-DX513 : Synology DX513 5 Bay Desktop Expansion	£346.16

= £350 + (5*£430) = £2,500

Build your own NAS Server

- Simpler to buy a commercial NAS server, but if you prefer you can also build the NAS server by yourself.

Brian's Blog Blog Archives About Me

FEB 3RD, 2016 | COMMENTS

DIY NAS: 2016 Edition

Final Parts List

Component	Part Name		Count	Cost
Motherboard	ASRock C2750D4I	specs	1	$418.04
Memory	Crucial 16GB Kit (8GBx2) DDR3 ECC	specs	1	$86.99
Case	U-NAS NSC-800 Server Chassis	specs	1	$199.00
Power Supply	Athena Power AP-U1ATX30A	specs	1	$43.14
SATA Cables	Monoprice 18-Inch SATA III 6.0 Gbps (Pkg of 5)	N/A	2	$6.99
OS Drive	SanDisk Cruzer 16GB USB Flash Drive	specs	1	$7.31
Cache Drives	Samsung 850 EVO 120GB SSD	specs	2	$67.99
Storage HDD 1	WD Red 4TB NAS - WD40EFRX	specs	3	$149.49
Storage HDD 2	Seagate NAS HDD 4TB (ST4000VN000)	specs	4	$139.00
			TOTAL:	$1,894.93

blog.brianmoses.net

ars TECHNICA UK SIGN IN

GEAR & GADGETS —

How to plan and build your own home NAS

There are lots of ins and outs, but you'll be fine if you begin with a solid plan.

ANDREW CUNNINGHAM (US) - 11/2/2016, 10:33

arstechnica.co.uk

Home Network NAS: Buy vs. Build?

By Joseph Moran

When you want to make storage available on a home network, two options usually come to mind. One is to simply set up shared folders on any PC, which is quick and doesn't cost anything but isn't centralized.

For centralized storage that isn't dependent on a single PC, a NAS device is a far better option. But buying a ready-made NAS device isn't the only way to get network storage. Here are some other options that may in some cases save money or provide more storage flexibility.

practicallynetworked.com

Home / Hardware

How to Build Your Own Network-Attached Storage System

10 COMMENTS

By Nate Ralph, PCWorld

Oct 16, 2011 6:00 PM

pcworld.com

BUILD GUIDE

FreeNAS Box (2016)

by Zenjiro

ZFS is the Primary Focus of this Build! - ZFS Primer

FreeNAS

The World's #1 Storage Operating System with over 8.5+ Million Downloads

Enterprise-Grade Features, Open Source, BSD Licensed

DOWNLOAD

What is FreeNAS?

FreeNAS is an operating system that can be installed on virtually any hardware platform to share data over a network. FreeNAS is the simplest way to create a centralized and easily accessible place for your data. Use FreeNAS with ZFS to protect, store, backup, all of your data. FreeNAS is used everywhere, for the home, small business, and the enterprise.

http://www.freenas.org/

NAS Server CUTTING COSTS

http://www.cnet.com/uk/how-to/raspberry-pi-as-cheap-nas-solution/

Networked hard drives are super convenient. You can access files no matter what computer you're on -- and even remotely.

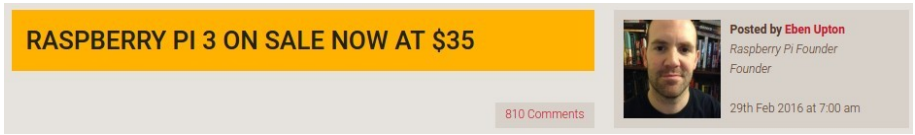
But they're expensive. Unless you use the Raspberry Pi.

If you happen to have a few of hard drives laying around you can put them to good use with a Raspberry Pi by creating your own, very cheap NAS setup. My current setup is two 4TB hard drives and one 128GB hard drive, connected to my network and accessible from anywhere using the Raspberry Pi.

For starters, you need an external storage drive, such as an HDD, SSD or a flash drive.

You also need a Raspberry Pi. Models 1 and 2 work just fine for this application but you will get a little better support from the **Raspberry Pi 3.** With the Pi 3, you're still limited to USB 2.0 and 100Mbps via Ethernet. However, I was able to power one external HDD with a Pi 3, while the Pi 2 Model B could not supply enough power to the same HDD.

In my Raspberry Pi NAS, I currently have one powered 4TB HDD, one non-powered 4TB HDD and a 128GB flash drive mounted without issue. To use a Pi 1 or 2 with this, you may want to consider using a powered USB hub for your external drives or using a HDD that requires external power.



https://www.raspberrypi.org/blog/raspberry-pi-3-on-sale/

http://www.bbc.co.uk/news/technology-35667990

NAS SERVERS #1

PCMAG.COM

PCMag UK | Guide

The 10 Best Network-Attached Storage (NAS) Devices of 2016

Networking

What's the best NAS drive for business?

Read our guide to find out what to look for when buying a NAS for business, and which NAS drives are best

Dave Mitchell 29 Dec 2014

CNET › Computer Accessories › Storage › Best hard drives and storage devices of 2016 ›

Storage:

Best network attached storage of 2016

Updated 6 June 2016 11:44 pm BST

NETWORKWORLD

SLIDESHOW

Best NAS boxes for small business

MORE

By James E. Gaskin, Network World | Oct 1, 2015 3:00 AM PT

NAS SERVERS #2

PCMag UK | Storage Devices - Reviews and Price Comparisons from PC Magazine | Feature

How to Buy Network-Attached Storage (NAS)

BY SAMARA LYNN 18 JUN 2014, 11:21 P.M.

Which NAS Is Right for You?

There are many varieties of use cases for NAS products. Luckily, there's a wide range of devices available—and many of them are configurable as well, which lets you further tailor a solution for your specific needs. Whether it's for home or a business, security, capacity, backup, and file compatibility should be key factors in determining which NAS you choose. The other features are mainly extras, which will be of greater or lesser importance depending on your particular needs.

Measuring NAS Performance

Like PCs, NAS units perform better with improved processors and increased memory. Similarly, the better the processor and the more installed memory, the higher the price. One of the fastest performing NASes we've tested is **ixSystems' FreeNAS Mini**. This device owes its superior performance to its Intel Core i3 processor and 8GB of RAM.

If you know your NAS will be handling a lot of I/O operations (such as users saving and retrieving high volumes of data on a regular basis) it pays to go with a NAS that has a nimble processor and to max out the memory. Most SMB NASes ship with Atom or Intel processors, while more inexpensive devices for home often use Marvell chips.

PCMag UK | Networking Reviews, Ratings & Comparisons | iXsystems FreeNAS Mini | Review

iXsystems FreeNAS Mini

EDITOR RATING: EXCELLENT (4.0)

REVIEW | COMMENTS | SPECS

PROS

Powerful hardware. Highly extensible. Feature-packed.

CONS

No dual Ethernet, by default. Complicated drive recovery. Steep learning curve for the non-Unix experienced.

BOTTOM LINE

FreeNAS Mini provides administrators the ultimate in control over their NAS, thanks to the extensibility of its open-source software, despite some exasperation with the Unix/Free BSD software in administering.

NAS Security

Security is always a concern, whether it's for home hardware or business networks. Many of the NAS devices we've reviewed support file encryption. Many also offer a **variety of security controls** to protect the NAS from intruders with firewall-like access protection. For example, business NAS devices often have physical security, such as locked enclosures or Kensington Security Locks (or K-Slots), which **tether the NAS to a wall or desk**. The QNAP TS-259 is one example of a NAS that has K-Slots on its chassis.

Finally, all NASes have user accounts and authentication methods requiring a username and password to access the device.

Clustered NAS "ADVANCED NAS"

ComputerWeekly.com

Clustered NAS vs traditional NAS solutions

By contrast, clustered NAS allows horizontal scaling across a number of devices with all of them being active and able to see all files in the cluster. This has a number of advantages:

- If your storage servers become CPU/memory-bound, you can add a device to gain processing power without adding disk.
- If you run out of storage, you can add disk that all devices can see, but you don't have to purchase additional devices.
- A device failure is non-disruptive, and the load of the failed unit can be spread across the whole cluster.

There's far less effort involved in managing clustered NAS compared with multiple traditional NAS devices, and I have found in discussions with colleagues in the industry that with clustered NAS we can manage in excess of 1 petabyte (PB) per full-time equivalent (FTE) employee.

SnapScale X2 Clustered NAS Storage by Overland Storage

Displaying products 1 - 6 of 6 results

SnapScale X2 by Overland Storage - Three Node Bundle 24TB NL-SAS Scalable Clustered NAS Part # OV-SSN301004

List Price: ~~$24,999.00~~
Our Price: Call for Pricing

Add to Cart | More Info

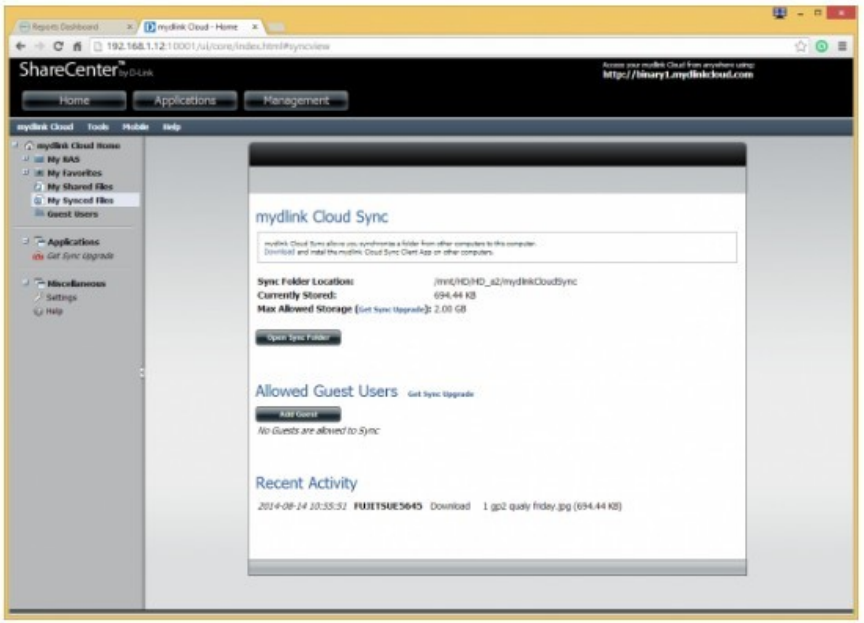
Isilon adds entry-level clustered NAS, drops pricing

by
Jo Maitland
TechTarget

Published: 21 Jan 2007

Moving out of its corner at the high end of the NAS market, Isilon introduces a more affordable version of its clustered NAS product for under $40K.

NAS CLOUD FROM NAS SERVER PROVIDERS



The appliances here can also provide Dropbox-like file-syncing services. D-Link offers its Cloud Sync app; Netgear has ReadyDROP; Qnap's version is called myQNAPcloud; and Synology offers Cloud Station.

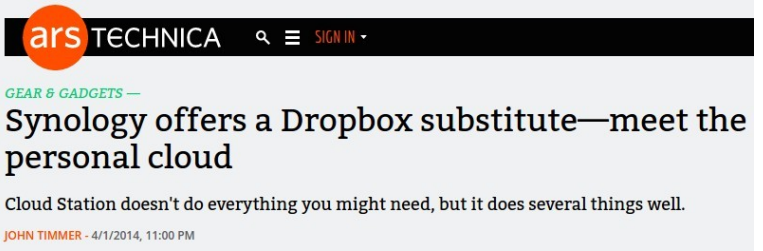
http://dilbert.com/strip/2016-04-04

- D-Link offers its ShareCenter Sync
- Netgear has ReadyDROP
- Qnap's version is called myQNAPcloud
- Synology offers Cloud Station

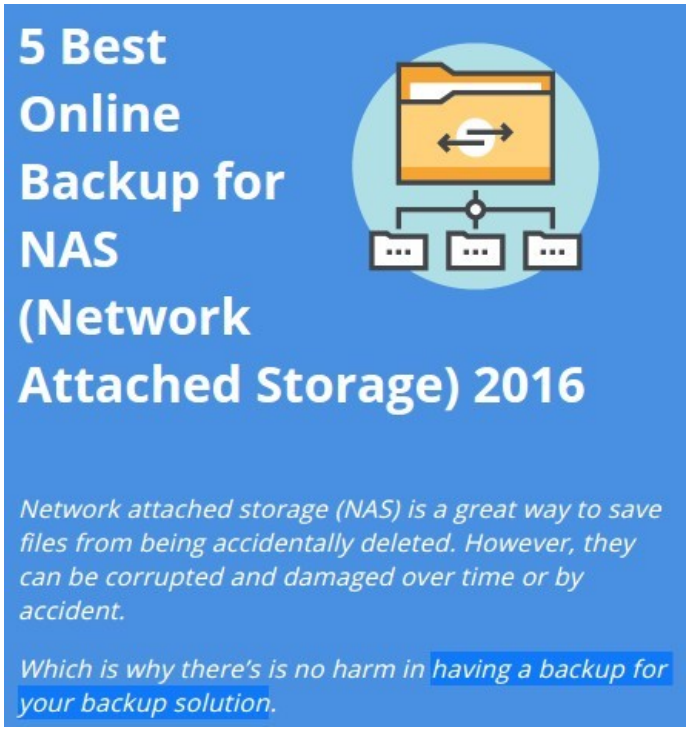
Sync files between Synology NAS and your computer using Cloud Station

Overview

Cloud Station for Synology NAS is a file-syncing application that lets you easily synchronize files on your Synology NAS with other devices, such as computers or mobile devices (with DS cloud). Install **Cloud Station Server** on your Synology NAS and **Cloud Station Drive** on your computer, to automatically sync files stored on your computer to your Synology NAS.



Additional Cloud Backup



Rang	Company Score	Price	Link
1	Winner CRASHPLAN www.crashplan.com ★★★★★	$ 5.99 PER MONTH Unlimited GB STORAGE All Plans	Visit Crashplan Review
2	IDrive www.idrive.com ★★★★★	$ 3.72 PER MONTH 1000 GB STORAGE All Plans	Visit IDrive Review
3	CARBONITE www.carbonite.com ★★★★½	$ 5.00 PER MONTH Unlimited GB STORAGE All Plans	Visit Carbonite Review
4	livedrive www.livedrive.com ★★★★½	$ 8.00 PER MONTH Unlimited GB STORAGE All Plans	Visit Livedrive Review
5	SOS www.sosonlinebackup.com ★★★★☆	$ 4.99 PER MONTH 50 GB STORAGE All Plans	Visit SOS Online Backup Review

http://www.carbonite.com
http://www.idrive.com
http://www.crashplan.com
http://www.backblaze.com
http://www.sosonlinebackup.com

Beyond 120 TB

HAVE SOME PROFESSIONAL DATA VENDOR TAKE CARE OF IT?
OR USE THE SEMI-DIY SOLUTIONS BY BACKBLAZE

Protecting petabytes: Best practices for big data backup

searchdatabackup.techtarget.com

by
Todd Erickson
Features Editor

How do you protect the massive data sets? Learn the best practices and products used for big data backup disaster recovery.

NAS4Free™
The Free Network Attached Storage Project

http://www.nas4free.org/

£7,270 for 180TB (£0.04 per GB)
backblaze.com

eRacks/NAS72

Our eRacks/NAS72 is a high-storage-density 4U rackmount storage server with 72 removable drive bays with up to **576TB of storage**.

openstack

Suitable for Cloud Storage, OpenStack, SDS (Software-defined Storage) with the leading Open Source SDS solutions, NAS/Local LAN usage, media libraries, or any number of other storage uses, the eRacks/NAS72is a truly petascale solution - ten eRacks/NAS72 servers in a standard 42u rack gives you **over 5 Petabytes of raw storage** - with 2U of room to spare for a UPS, KVM, Network switch/firewall, or other of eRacks' rack accessories.

ceph - Your choice of Cloud software is available, including OpenStack, CloudStack, Eucalyptus, OpenNebula, ProxMox, or other - just specify in the "Notes" field when you place your order or request a quote.

- Your choice of Storage or SDS software, to fit inro a larger architecture or pre-existing cloud infrastructure - choices include Ceph, GlusterFS, MooseFS (an eRacks Partner), LizardFS (Also ans eRacks Partner), or others
- Your choice of NAS software / OS is available on request, including Samba, FreeNAS, NAS4Free, OpenMediaVault, or your choice - just specify it in the "Notes" field when you place your order.

SAMBA - Or, you may choose one of our regular distros, and/or simply ask for your preferred OS, packages, and storage options - Arch Linux, LizardFS,Ceph, MooseFS,Samba,NFS, Novell, Apple and any other available configuration requests - just add it in the Notes.

This model is available with single and dual multicore Xeon e5 v3-series CPUs.

Starting at **$21995**

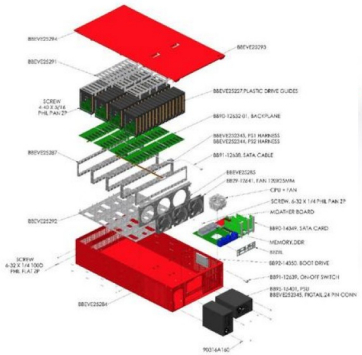
https://eracks.com/products/storage-servers-nas/

Backblaze Pod 6.0

Building your own server? Backblaze's DIY Pod 6.0 can be scaled to house 480TB of data thenextweb.com

by NATE SWANNER — 3 months ago in DESIGN & DEV

Building your own server 'pod' isn't exceedingly hard, but Backblaze just made it far easier with its new Pod 6.0 scheme, which can house up to 480TB of memory.



backuppods.com

How Built	Total Cost	Description
Backblaze	$8,733.73	The cost for Backblaze given that we purchase 500+ Storage Pods and 20,000+ hard drives per year. This includes materials, assembly, and testing.
You Build It	$10,398.57	The cost for you to build one Storage Pod 6.0 server by buying the parts and assembling it yourself.
You Buy It	$12,849.40	The cost for you to purchase one already assembled Storage Pod 6.0 server from a third-party supplier and then purchase and install 4TB hard drives yourself.

180TB Storage Pod 6.0 storage server with 4TB hard drives

	Storage Pod Version						
	1.0	2.0	3.0	4.0	4.5	5.0	6.0
Backblaze Cost ($)	7,867	7,394	7,568	9,305	8,688	7,974	8,734
Drive Size (TB)	1.5	3.0	3.0	4.0	4.0	4.0	4.0
Total Storage (TB)	67.5	135	135	180	180	180	240
Cost per GB	0.117	0.055	0.056	0.052	0.048	0.044	0.036

Storage Pod 6.0 Cost per GB for different drives

Drive Size	MFG	Model	Unit Price	Drive Cost	Pod Capacity	Pod Price	Cost per GB
4TB	Seagate	ST4000DM000(*)	114.99	6,899.40	240	10,364.07	0.043
6TB	Seagate	ST6000DM001(*)	214.00	12,840.00	360	16,304.67	0.045
4TB	HGST	HMS5C4040BLE640(*)	128.87	7,732.20	240	11,196.87	0.047
8TB	Seagate	ST8000DM002	318.85	19,131.00	480	22,595.67	0.047
8TB	WD (Red)	WD80EFZX	340.35	20,421.00	480	23,885.67	0.050
6TB	WD (Red)	WD60EFRX (*)	242.01	14,520.60	360	17,985.27	0.050
5TB	WD (Red)	WD50EFRX	192.99	11,579.40	300	15,044.07	0.050
5TB	Seagate	ST5000DM002	199.00	11,940.00	300	15,404.67	0.051
4TB	WD (Red)	WD40EFRX (*)	149.99	8,999.40	240	12,464.07	0.052
8TB	HGST Helium	HUH728080ALE600 (*)	418.70	25,122.00	480	28,568.67	0.060
6TB	HGST Helium	HUH728060ALE600	369.21	22,152.60	360	25,617.27	0.071

https://www.backblaze.com/blog/open-source-data-storage-server/

BACKBLAZE

SSD Data center WELL ONE DAY THEN

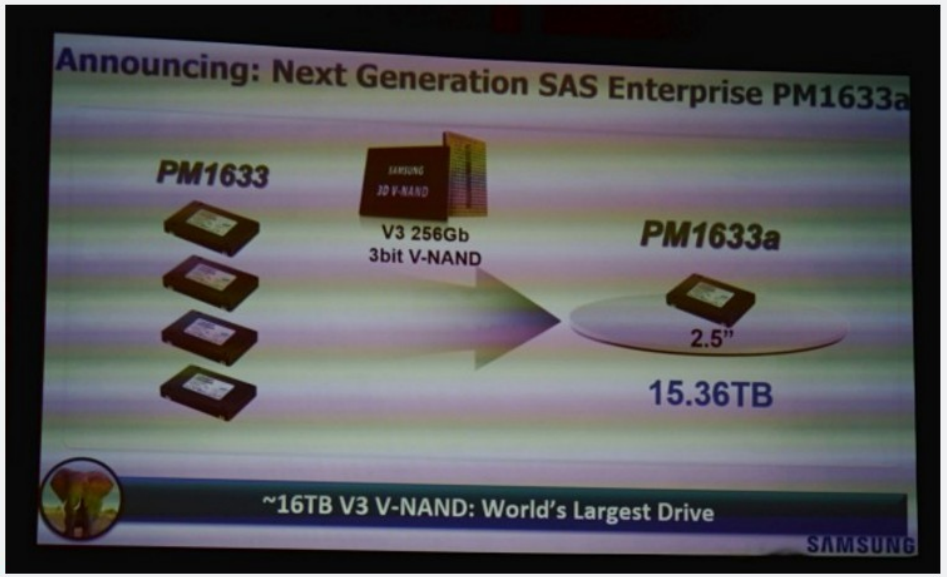
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE FORUMS

CRAZYSEXYCOOL —

Samsung unveils 2.5-inch 16TB SSD: The world's largest hard drive

Third-generation 3D V-NAND is now up to 48 TLC layers and 256Gbit per die.

SEBASTIAN ANTHONY (UK) - 8/13/2015, 2:14 PM



http://arstechnica.com/gadgets/2015/08/

ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE FORUMS

Seagate's 60TB SSD comes a year after Samsung's 15TB SSD.

SEBASTIAN ANTHONY (UK) - 8/11/2016, 1:46 PM

http://arstechnica.com/gadgets/2016/08/
3.5" form factor.

So as the Backblaze Pod 6.0 had 60 slots for 3.5 HDDs, you could build really fast ~3.6 petabyte NAS server (60 x 60 TB) with these.

References

- **A Full Hardware Guide to Deep Learning** (2015) by Tim Dettmers
- **Which GPU(s) to Get for Deep Learning: My Experience and Advice for Using GPUs in Deep Learning** (2014) by Tim Dettmers
- **Building a Deep Learning (Dream) Machine** (Sept 2015) by Roelof Pieters
- **Deep Learning GPU-Based Hardware Platform** Hardware and Software Criteria and Selection. By Mourad Bouache and John Glover | ICS-2016, Istanbul, Turkey May 31st 2016.
- **Build Personal Deep Learning Rig**: GTX 1080 + Ubuntu 16.04 + CUDA 8.0RC + CuDnn 7 + Tensorflow/Mxnet/Caffe/Darknet by Guanghan Ning
- Dmytro Prylipko's article on building a **"DIY: Deep Learning DevBox"**
- **Hardware Guide: Neural Networks on GPUs** (Updated 2016-1-30) by Joseph Redon