Scaling mobile GPUs to 1000 GFLOPS

On the 20th of April 2013 there was an interesting discussion between Jan Gray and David Kanter. Jan is a specialist in C++ and FPGAs (twitter, homepage). David is a specialist in CPU and GPU architectures (twitter, homepage). Both know their ways well in the field of semiconductors. It is always a joy to follow their short discussions when they happen, but there was something about this one that made me want to share it with special attention.
OpenCL on ARM: Growth-expectation of GFLOPS/Watt of mobile GPUs exceeds Moore's law. That's incredible!

Jan Gray: .@OpenCLonARM GFLOPS/W more a factor of almost-over Dennard Scaling. But plenty of waste still to quash. http://www.fpgacpu.org/papers/Gray_AutumnOfMooresLaw_SingularityUniversity_11-06-23.pdf ?

Jan Gray?: .@openclonarm Scratch Dennard tweet: reduced capacitance of yet smaller devices shd improve GFLOPS/W even as we approach end of Vdd scaling.

David Kanter: @jangray @OpenCLonARM I think some companies would argue Vdd scaling isn't dead...

Jan Gray: @TheKanter @openclonarm it's not dead, but slowing, we've gone from 5V to 1V (25x power savings) and have maybe several hundred mVs to go.

David Kanter: @jangray I reckon we have at least 400mV, so ~2X; slower than ideal, but still significant

Jan Gray: @TheKanter We agree, I think.

David Kanter: @jangray I suspect that if GPU scaling > Moore's Law then they are just spending more area or power; like discrete GPUs in the last decade

David Kanter: @jangray also, most positive comment I've heard from industry folks on mobile GPU software and drivers is "catastrophically terrible"

Jan Gray: @TheKanter Many ways to reduce power, soup to nuts. For ex HMC DRAM on interposer for lower energy signaling. I'm sure many tricks to come.
In a nutshell, all the reasons they think mobile GPUs can outpace Moore's law while staying under a certain power-usage.

It needs some background-info, so let's start the background of the first tweet, and then explain what has been said.

More than Moore for mobile GPUs

Requirements for Mobile GPUs were not really high until the iPhone and Android came along. Of course there was demand, but not from such a big market. This caused a big push forward of mobile GPUs. The reason it can grow faster than other processors is because catching-up is possible... at least from what I picked up.

For example Imagination Technologies promises a lot with PowerVR series 6:

"With a growing range of cores optimised for either area (maximising GFLOPS/mm2) or performance (maximising GFLOPS/mW), PowerVR Series6 GPUs can deliver 20x or more of the performance of current generation GPU cores targeting comparable markets. This is enabled by an architecture that is around 5x more efficient than previous generations. PowerVR Series6 GPU cores are designed to offer computing performance exceeding 100GFLOPS (gigaFLOPS) and reaching the TFLOPS (teraFLOPS) range enabling high-level graphics performance from mobile through to high-end compute and graphics solutions".

Currently, there are around 70 GFLOPS; and 1 TFLOPS should be reached before 2020, when they have the same pace in advancements as desktop GPUs. But they claim the reach it with Series 6!

What do Jan and David have to say about it?

Dennard's scaling

Jan takes back this reason a tweet later, but it's an interesting rule anyway:

Dennard's scaling rules observe that voltage and current should be proportional to the linear dimensions of a transistor, implying that power consumption (the product of voltage and current) will be proportional to the area of a transistor. This property implies that shrunk MOSFETs will consume less power, and forms the basis of Moore's Law. (source)
The projected end with current technologies (Silicon) is between 10 and 7 nm. Check our post Molybdenite and graphene to the helping hand for what can come as an alternative.

Current mobile GPUs are baked on 32 and even 45 nm, but will soon go under 25 nm. Nevertheless, this does not give the mobile GPUs an advantage over desktop GPUs.

Voltage scaling

Jan talks about "reduced capacitance of yet smaller devices" and Vdd. Smaller chip-sizes make it possible to work with less volts. Vdd is the voltage hooked up to the chip from outside. As explained in the tweets, this has gone down over the years.

(You can check out this PDF of OCW-course "Microelectronic Devices and Circuits" about scaling to learn more.)

Voltage can go down for other processors too. So not a strong reason why Imagination would go from 70 to 1000 GFLOPS, while desktop GPUs go from around 4000 to around 4000. But as mobile processors tend to be smaller, there is some advantage.

A side-note: another explanation is that the power-usage maximum will be loosened. Then the 1 TFLOPS GPU would not give maximum performance when on battery power.

Area Scaling

Years ago ARM's chips were under 10 mm2 (!), GPUs larger. Intel's CPUs have been over 100 mm2 for years, even with shrinking die-size.

Mobile chips therefore had an advantage over X86, but now they are getting less advantage with die-sizes over 100 mm2. See below for what CPU+GPU in the X86-world take for an area.

Trinity Physical Comparison (source: Anandtech)

Manufacturing Process
Die Size
Transistor Count

AMD Llano
32 nm
228 mm2
1.178B

AMD Trinity
32 nm
246 mm2
1.303B

Intel Sandy Bridge (4C)
32 nm
216 mm2
1.16B

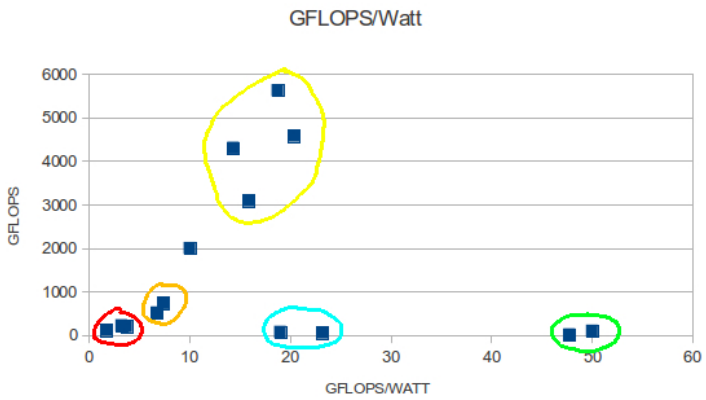
Intel Ivy Bridge (4C)
22nm
160mm2
1.4B

Seems that here's the stretchiest of all. But it also has its limits due to power-usage. Given the fact that larger chips use more power, it still has its limits.

"Many tricks to come"

Jan mentions tricks are also helping, mentioning a memory-transfer optimisation. ARM and Imagination are known to have a lot of expertise and IP in SOC-optimisations, low-power and memory-transfers; something with which AMD and Intel have just started.

In the below graph (taken from the article "Processors that can do 20+ GFLOPS per Watt") you see the desktop GPUs (yellow) and the mobile processors (light-blue). The empty area at the upper-right is where most interest is. Could they end up right of the desktop GPUs? The question is still how big that advantage can be.



One of the reasons to co-organise the LEAP-conference is exactly to get answers to that one question. Why would vendors of low-power processors outpace both Moore's law and desktop GPUs?

Jan and David shared their opinion, but I would really like to hear form the manufacturers themselves.

You are very welcome to give your own views in the comments!

Also, if you have found interesting articles (partly) answering this question, please share them with the other readers...

You can join the discussion on Twitter about mobile GPUs reaching 1 TFLOPS at #terascale.]