Towards a Methodology for Creating Time-critical, Cloud-based CUDA Applications

Louise Knight
Polona Štefanič
Matej Cigale
Andrew Jones
Ian Taylor

School of Computer Science & Informatics, Cardiff University, UK

IT4RIs 18/01/18

Some context

- I joined SWITCH in late October 2017
- My main task has been to explore the potential for extending SWITCH to support development and operation of time-critical cloud-based CUDA applications
- The survey presented here was performed in order to understand the “landscape” of time-critical CUDA applications and relate them to the requirements and performance parameters for my own CUDA application (algorithms for detecting co-evolution in protein sequence data)

Some context (cont'd)

- Practical experiments are being undertaken (outside the scope of this presentation) to explore the performance of these co-evolution algorithms on Amazon Web Services GPU-backed instances while varying different parameters, in order to obtain results that should inform cloud deployment of applications such as those surveyed here
- These results, in turn, have the potential to inform the enhancement of the SWITCH planning, provisioning and self-adaptation algorithms to support CUDA applications

What is CUDA?

- NVIDIA graphics cards for parallel processing
- Relatively low-cost over long-term use (in comparison, to e.g. buying time on a supercomputer)
- CUDA programming language derived from C (cudaMemcpy)
- SIMD - Single Instruction Multiple Data

Louise Knight KnightL2@cardiff.ac.uk

Survey of time-critical CUDA applications

Four categories:

1. Environment-related
2. People/face detection
3. Medical applications
4. Materials-related

(Miscellaneous)

Also, image processing/non-image processing

Louise Knight KnightL2@cardiff.ac.uk

QoS metrics

- **Time:** runtime, time per unit of computation, processing speed, processing rate, communication latency, latency per frame
- **Quality of results:** accuracy, correlation coefficient, average absolute difference, sensitivity, false positives count, peak signal-to-noise ratio, mean-square-error
- **Data:** throughput, memory access
- **Compute:** floating point operations count
- **Costs:** performance-per-watt, performance-per-dollar, implementation cost

Taxonomy

- Many terms for same thing, even within fields, e.g. runtime
  - Convergence time
  - Detection time
  - Processing speed
  - Reconstruction time
- Image processing versus non-image processing (quality)

Louise Knight KnightL2@cardiff.ac.uk

Metrics with most influence

- **Environment-related:**
  - Accuracy (correlation coefficient) - focus on this related to critical nature of applications
- **People/face detection:**
  - Accuracy
  - Sensitivity
  - False positives count
  - More “serious” applications (surveillance systems, pedestrian detection) to less “serious” (virtual reality)

Louise Knight KnightL2@cardiff.ac.uk

Metrics with most influence

- **Medical applications:**
  - Accuracy (peak signal-to-noise ratio, mean-square-error) - intra-operative applications
  - Throughput - work within constraints
- **Materials-related:**
  - Runtime only
- **Miscellaneous:**
  - Accuracy
  - Throughput
  - Memory access
  - Floating point operations

Louise Knight KnightL2@cardiff.ac.uk

How does this relate to SWITCH?

- SWITCH motivated by QoE as well as QoS
- QoS allows to measure user's experience of application and adapt as necessary
- Determine infrastructure needed to run CUDA problems, facilitate analysis in a specific time (real-time)
- Some use cases in Miscellaneous category similar to two SWITCH use cases:
    - Synchronising multiple video streams; enhanced audience experience during live sporting events - similar to MOG
    - Immersive 3D video-conferencing - similar to WT

Louise Knight KnightL2@cardiff.ac.uk

Possible SWITCH extensions using CUDA

- Amazon Web Services (AWS)
- Adapt SIDE/DRIP in future?
- Consistency of time within which result returned as important as speed
- Evaluate application for how well support CUDA
- New class of applications that require fast response time with limited investments - e.g. Modelling system of Stefanic et al. - larger example sets analysed in process of deriving model; produce better models faster

Louise Knight KnightL2@cardiff.ac.uk

Future work

- Begun experiments with AWS
- Multiple, single-GPU instances
- Single geographical Region (for now)
- Changing parameters to see influence on QoS
- Insight into how SWITCH may support CUDA in the future

Louise Knight KnightL2@cardiff.ac.uk

Thank you for listening

References

P. Stefanic, M. Cigale, A. Jones, and V. Stankovski, “Quality of service models for microservices and their integration into the switch ide,” in *2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W)*, Sept 2017, pp. 215–218.

M. Mehrabi, S. Lafond, and L. Wang, “Frame Synchronization of Live Video Streams Using Visible Light Communication,” *2015 IEEE International Symposium on Multimedia (ISM)*, pp. 128–131, 2015.

J. Maillard, M. Leny, and H. Diakhate, “Enhancing the audience experience during sport events: Real-time processing of multiple stereoscopic cameras,” *Annales des Telecommunications/Annals of Telecommunications*, vol. 68, no. 11-12, pp. 657–671, 2013.

I. Feldmann, W. Waizenegger, N. Atzpadin, and O. Schreer, “Realtime depth estimation for immersive 3D videoconferencing,” *3DTV-CON 2010: The True Vision - Capture, Transmission and Display of 3D Video*, 2010.

Louise Knight KnightL2@cardiff.ac.uk