Università
della
Svizzera
italiana

Faculty
of Informatics

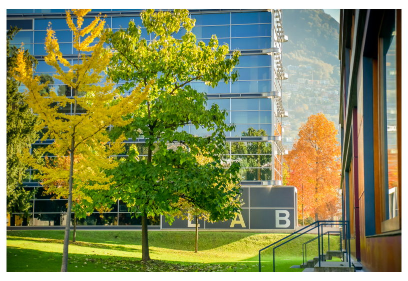
Institute of
Computational
Science
ICS

Large Scale Xeon Phi Parallelization of a Deep Learning Language Model

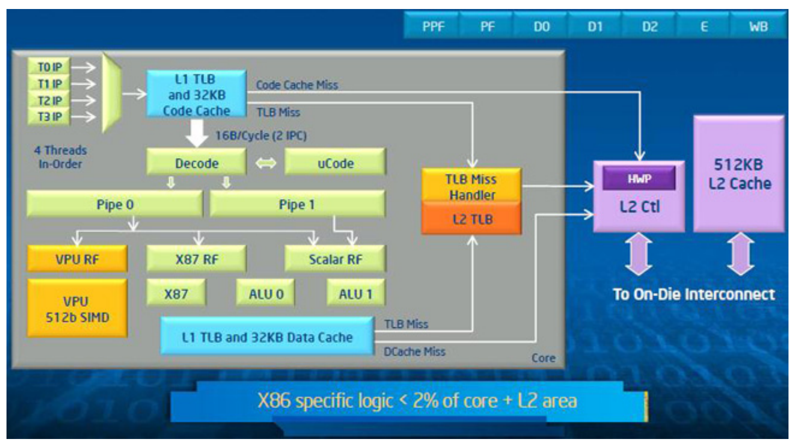
Tim Dettmers, Hanieh Soleimani, Olaf Schenk

Introduction to Software Atelier Course

The Software Atelier master course at the Computational Science institute of the University of Lugano focused on the supercomputing and simulation. It was presenting advanced topics in parallel computing and numerical simulation for prospective computational and software engineers. The Software Atelier master course at the Computational Science institute of the University of Lugano focused on the supercomputing and simulation.

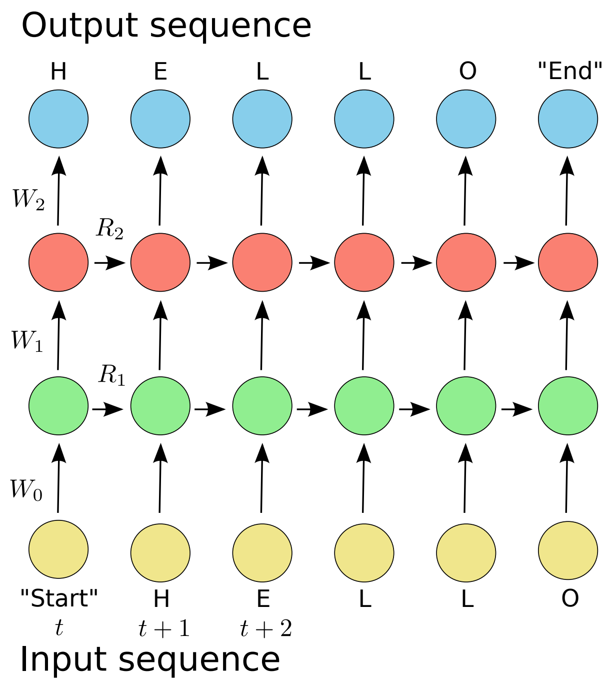


Intel Xeon Phi & Salomon Cluster Overview



Salomon Cluster has 1008 nodes, 576 of which have Xeon Phi 7120P accelerators with 2.4 TFLOPS for single precession and 352GB/s memory bandwidth. Also It has a FDR56 Infiniband interconnect which is connected in a 7D enhanced hypercube architecture

Deep Learning Language Model



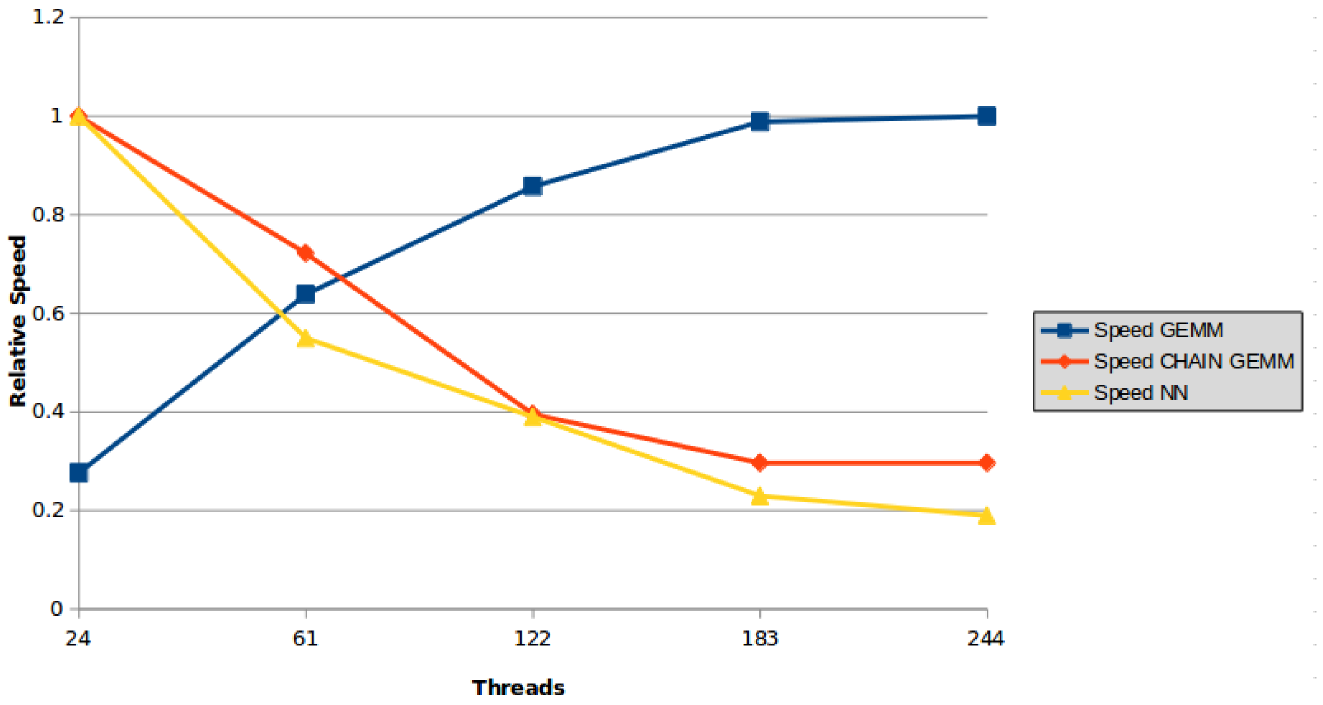
$$\mathbf{Y_{i+1}^{t+1}} = \mathbf{A_i^t W_i} + \mathbf{A_{i+1}^{t-1} R_i} + \mathbf{B_i} \tag{1}$$

$$\mathbf{A_{i+1}^{t+1}} = \sigma(\mathbf{Y_{i+1}^{t+1}}) \tag{2}$$

$$\frac{\partial \mathbf{E}}{\partial \mathbf{W}} = \frac{\partial \mathbf{E}}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{W}} = \frac{\partial \mathbf{E}}{\partial \mathbf{A_{i+1}}} \frac{\partial \mathbf{A_{i+1}}}{\mathbf{Y}} \frac{\partial \mathbf{y}}{\partial \mathbf{W}} \tag{3}$$

Here (1) and (2) are for the forward pass, (2) is the gradient for the backward pass.

Results



For chained small general matrix multiplications we have:

- (GEMM) is fast if many threads are used (about 70% peak performance) and slow if few threads are used (25% of peak performance)
- GEMM is slow if many threads are used and the successive matrix dimensions differ (typical for a neural network; 10% of peak performance) and faster for few threads (20% of peak performance)
- It follows that neither many, nor few threads are efficient for feedforward and recurrent neural networks
- Random number generation on Xeon Phi for successive matrices of small size (rough size is 128x1200) is slow independent of the number of threads that are used. This decreases neural network performance by a factor of 10

Discussion

- It is unclear what exactly causes the performance decreases for matrix operations (GEMM and random number generation) for successive matrices with different dimensions, but it has somehow to do how threads are scheduled and how resources are shared among threads.
- These problems made it impossible to even train a simple deep learning language model on Xeon Phi accelerators in a timely manner; the training time on one Xeon Phi would exceed one year

Future work

- Bottlenecks in GEMM and random number generation should be analyzed more carefully on both the software and hardware level
- Once problems with GEMM and random number generation are solved one can parallelize the deep learning language model on multiple nodes using MPI

References

- Strom, N. (2015). Scalable Distributed DNN Training Using Commodity GPU Cloud Computing. Interspeech2015.
- Dettmers, T. (2016). 8-bit Approximation for Parallelism in Deep Learning. ICLR 2016.