MANCHESTER 1824
The University of Manchester

Microstructural Analysis of Bone Using the Knights Landing Processor

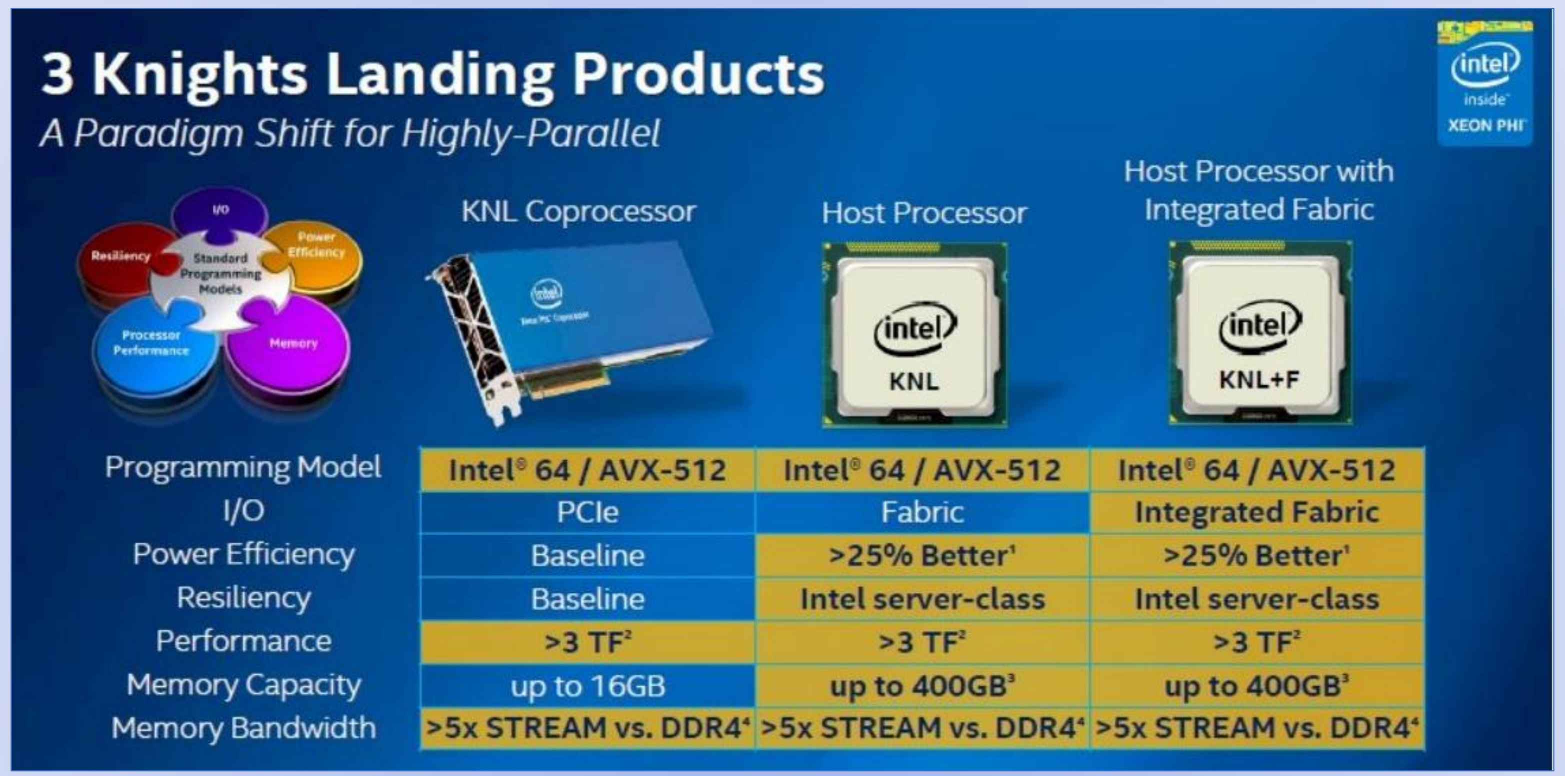
Name: Li Juan Chan

Project Supervisor: Dr. Lee Margetts

1. Aim

- To determine whether the Knights Landing (KNL) processor is fast enough to analyse bone in a clinical setting.
- Reducing the time taken to compute the finite element models will allow more patients to benefit from this capability.

2. Overview of KNL



	KNL Coprocessor	Host Processor	Host Processor with Integrated Fabric
Programming Model	Intel® 64 / AVX-512	Intel® 64 / AVX-512	Intel® 64 / AVX-512
I/O	PCIe	Fabric	Integrated Fabric
Power Efficiency	Baseline	>25% Better[1]	>25% Better[1]
Resiliency	Baseline	Intel server-class	Intel server-class
Performance	>3 TF[2]	>3 TF[2]	>3 TF[2]
Memory Capacity	up to 16GB	up to 400GB[3]	up to 400GB[3]
Memory Bandwidth	>5x STREAM vs. DDR4[4]	>5x STREAM vs. DDR4[4]	>5x STREAM vs. DDR4[4]

Figure 1: Variants of KNL products (Source: Codreanu, Rodriguez and Saastad, 2017)

Specification
- 64 processor cores, each with the speed of 1.3 GHz.
- 4 hyperthreads per processor core.
- High capacity DDR4 memory of 96 GB.
- High bandwidth MCDRAM of 16 GB.
- 512-bit SIMD instruction with each core operates vector of size 8 per clock cycle.
- Three types of memory modes: Cache mode, Flat mode and Hybrid mode.

3. Comparison between Xeon and KNL

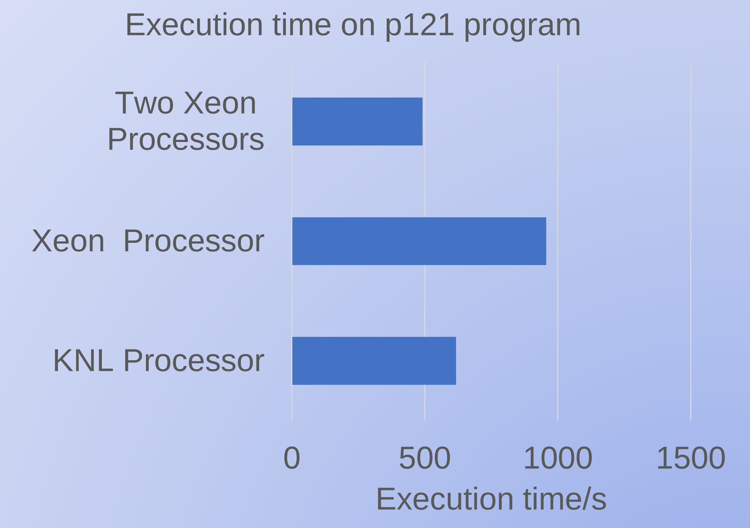


Figure 2 : Total execution time of Xeon and KNL on ParaFEM

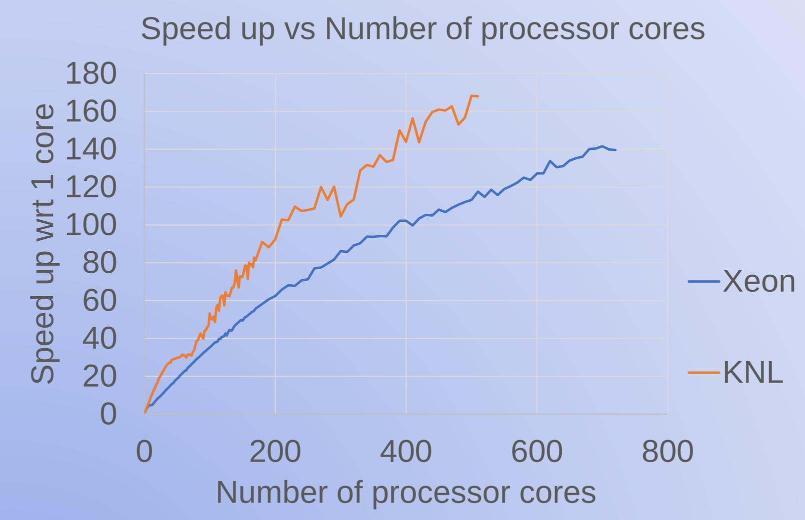


Figure 3: Speed up of Xeon and KNL on ParaFEM

- The performance of the Xeon and the KNL processors are compared using the code from an open-source finite element software, ParaFEM (Smith, Griffiths and Margetts, 2013).
- One KNL processor performs better than one Xeon processor, but worse than two Xeon processors on ParaFEM.
- The clock speeds of a KNL processor with one hyperthread and a Xeon processor are 83.2 GHz and 32.4 GHz respectively.
- Theoretically, one KNL processor should perform better than one and two Xeon processors.
- The factors that cause the difference between the theoretical and the experimental performance may include the parallel overhead and the parallelism of the code.
- The parallel overhead is caused by the increase of the amount of time needed to coordinate the parallel tasks as the number of processor cores increase.
- Additionally, the socket of the KNL processor is smaller than that of the Xeon processor, making the KNL processor require less power and cooling.
- The speed up with respect to one core is the ratio of analysis time of one processor core to the total number of processor cores used in parallel execution.
- The speed up of the KNL processor is better than that of the Xeon processor.
- Due to the less advanced processor cores in the KNL, the benefit obtained from executing the software in parallel is more significant in the KNL processor than the Xeon processor.

4. Optimisation of Code using OpenMP

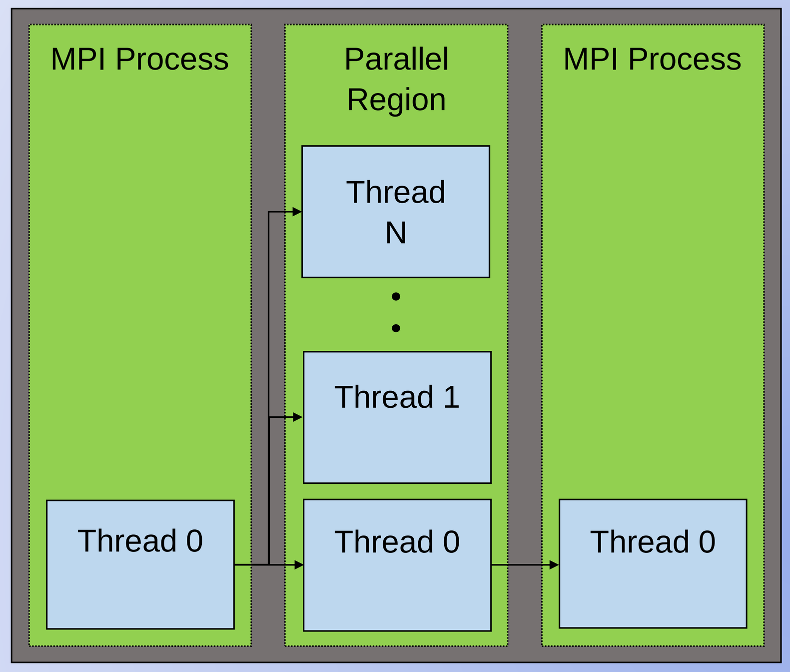


Figure 4: OpenMP fork join model

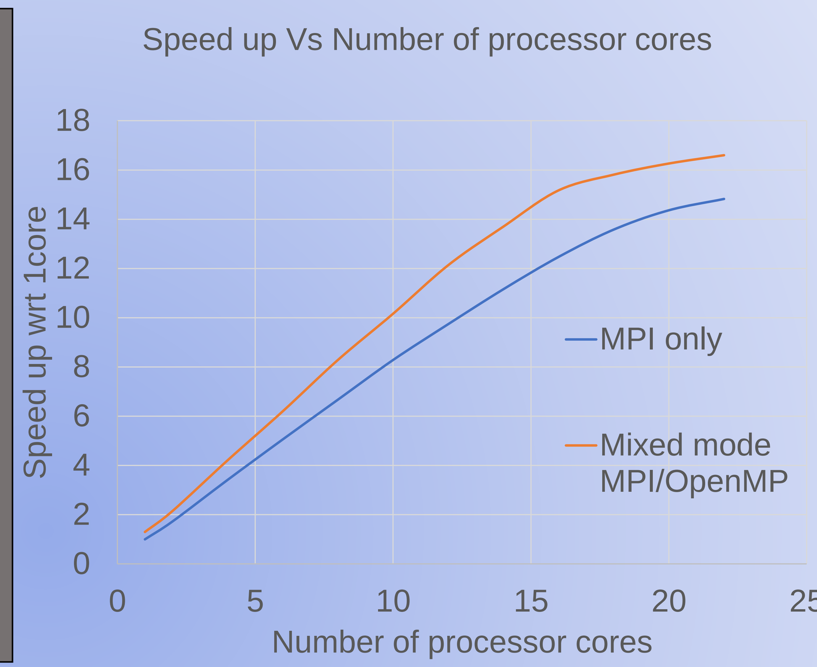


Figure 5: Speed up of MPI and Mixed MPI/OpenMP

- OpenMP is a standardised Application Program Interface (API) for shared memory system.
- OpenMP consists of three main components, which are compiler directives, runtime library routines and environment variables.
- OpenMP directives are included in the source code to advise the compiler that the program can be run using the hyperthreading feature of the KNL processor.
- Using both MPI processes and OpenMP threads reduces the analysis time compared to using MPI processes only as shown in Figure 5.
- Even though the improvement may not be significant, the benefit is still essential as the optimisation using OpenMP can be done in no time.

5. Benchmark Analysis

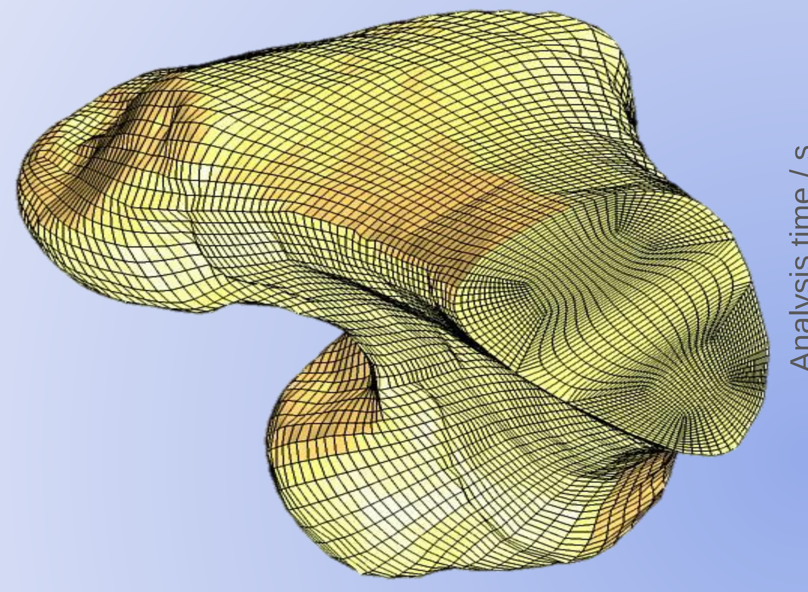


Figure 6: Unstructured mesh (Source: TrueGrid, accessed 1 Dec 2018)

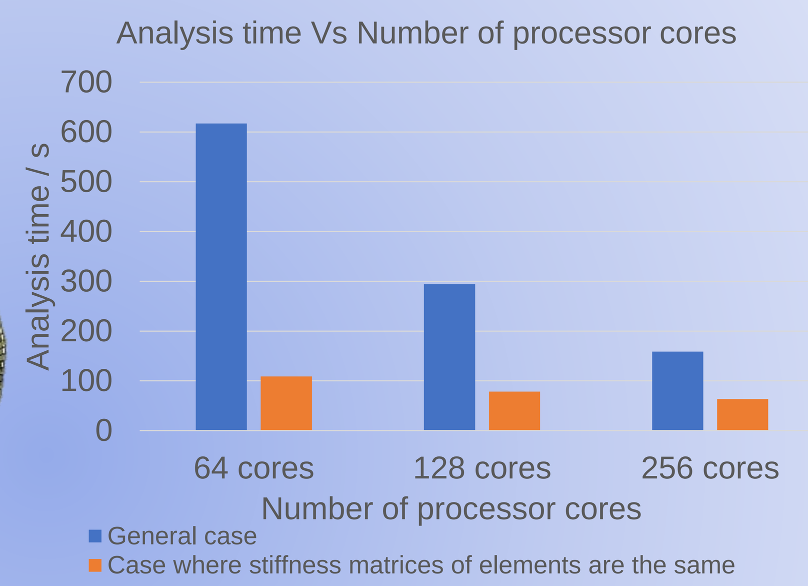


Figure 7: Comparison of analysis time for different cases

- The bar chart shows that the analysis time for the structured mesh is significantly lower than that of the unstructured mesh.
- An example of an unstructured mesh is shown in figure 3 in which the stiffness matrices of all the elements may be different.
- The structured mesh is the mesh in which the stiffness matrices are the same.
- The structured mesh can be used in bone analysis as the voxels in the bone image have the same geometry and may be assumed to have the same material properties.

6. Conversion of CT scan into Finite Element

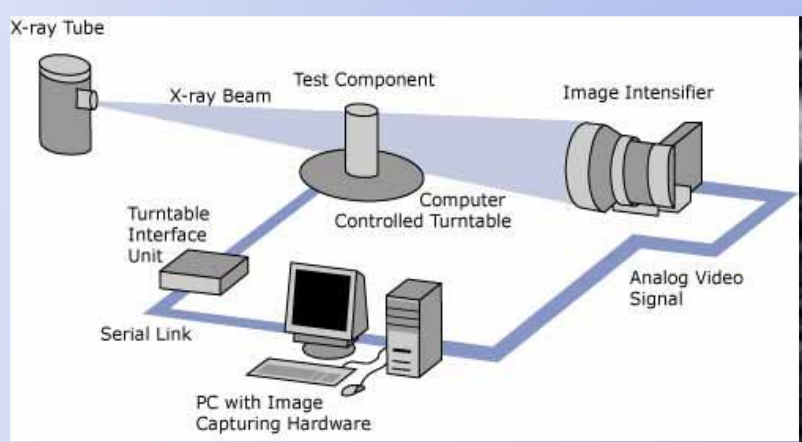


Figure 8: Materials CT scanning (NDT Resource Center, accessed 1 Dec 2018)

Figure 9: Bone image obtained from CT tomography (Source: Figure provided through personal communication with Alessandro Melis from the University of Sheffield)

- The bone is scanned in 360° to obtain multiple bone images similar to Figure 9.
- The bone images are sent to a computer to reconstruct the images into 3D model.
- The voxels in the 3D model are converted to hexahedral elements and are analysed using the Finite Element Method.

7. Further Work

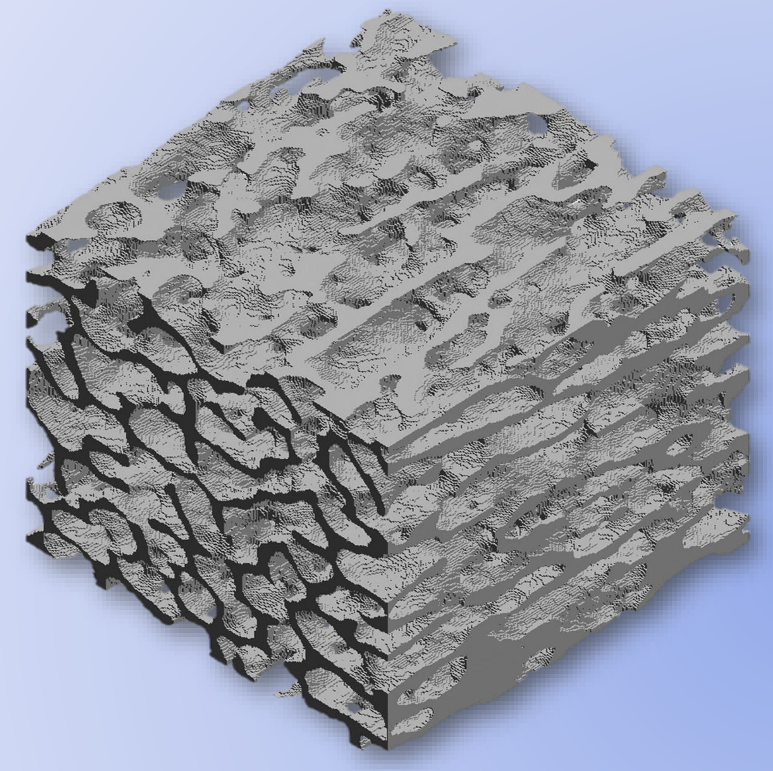


Figure 10: X-ray tomography picture of healthy trabecular bone (Source: Levrero-Florencio *et al.*, 2016)

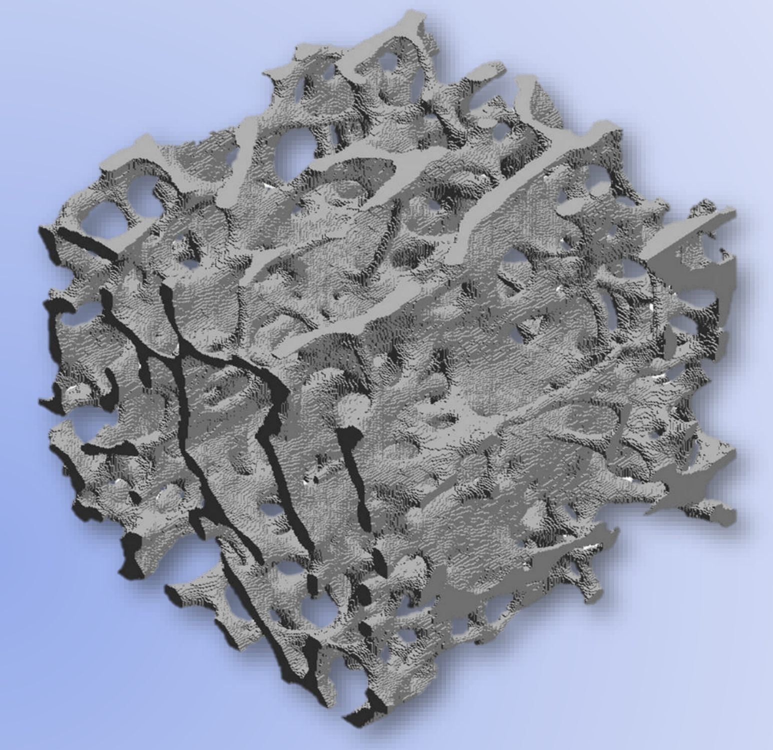


Figure 11: X-ray tomography picture of diseased trabecular bone (Source: Levrero-Florencio *et al.*, 2016)

- Figure 10 shows the picture of a healthy trabecular bone whereas figure 11 shows the picture of a low density diseased trabecular bone.
- The voxels in the bone image are converted to hexahedral elements using the MATLAB script written by the University of Sheffield.
- It provides an input deck for ParaFem, run in ARCHER and the KNL processor.
- The output of this project is a fast computer program for analysing bone model for the KNL processor. This will be used by UK researchers, particularly in Manchester, Edinburgh and Sheffield.

8. Potential Hardware for Further Evaluation

Figure 12: Cavium's ThunderX2 (Source: Kennedy, 2017)

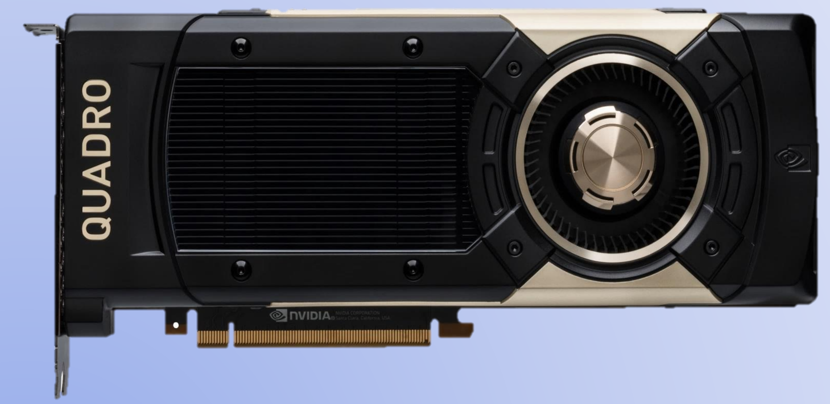


Figure 13: NVIDIA Quadro Volta GV100 (Source: Pette, 2018)

- The Cavium ThunderX2 processor is based on ARMv8 architecture and is integrated into the ARM-based supercomputer in the UK, with the code name of 'Isambard'.
- It is comparable to the KNL processor because ARM architecture processor is known for its high energy efficiency.
- The KNL processor inherits some of the characteristics of GPUs, in which it consists of many low clock speed processing units.
- GPUs may contain hundreds to thousands of processor cores, while KNL processor contains only tens of processor cores.
- The existence of latency of data transferring between host processors and GPUs may limit the performance of GPUs.
- However, GPUs are a lot cheaper than the KNL processor and may perform better than the KNL processor in embarrassingly parallel software.
- NVIDIA Quadro Volta GV100 is a GPU, which is specially designed to give superior performance in deep learning and is used in the fastest supercomputer in the world, 'Summit' (Top500, 2018).

9. References

Codreanu, V., Rodriguez, J. and Saastad, O. W. (2017). *Best Practice Guide – Knights Landing, January 2017.* Available at: http://www.prace-ri.eu/best-practice-guide-knights-landing-january-2017/, (Accessed: 1 Dec 2018).

Kennedy, P. (2017). *Cavium ThunderX2 Dual Socket System Spotted Live.* Available at: https://www.servethehome.com/cavium-thunderx2-dual-socket-system-spotted-live/, (Accessed: 1 Dec 2018).

Levrero-Florencio, F., Margetts, L., Sales, E., Xie, S., Manda, K. and Pankaj, P. (2016). 'Evaluating the macroscopic yield behaviour of trabecular bone using a nonlinear homogenisation approach', *Journal of the Mechanical Behavior of Biomedical Materials*, 61, pp. 384-396, [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S1751616116300728 (Accessed: 1 Dec 2018).

NDT Resource Center. (no date). *Computed Tomography*. Available at: https://www.nde-ed.org/EducationResources/CommunityCollege/Radiography/AdvancedTechniques/computedtomography.htm, (Accessed: 1 Dec 2018).

Pette, B. (2018). *NVIDIA Transforms the Workstation for the Age of Deep Learning.* Available at: https://blogs.nvidia.com/blog/2018/03/27/quadro-gv100-deep-learning-simulation/, (Accessed: 1 Dec 2018).

Smith, I. M., Griffiths, D. V. and Margetts, L. (2013). *Programming the Finite Element Method*. 5th edn. West Sussex: Wiley & Sons Ltd.

Top500. (2018). *November 2018.* Available at: https://www.top500.org/lists/2018/11/, (Accessed: 1 Dec 2018).

TrueGrid. (no date). *Femur.* Available at: http://www.truegrid.com/bone3.html, (Accessed: 1 Dec 2018).