Whitepaper

DigiCortex™

Large-Scale Biological Neural Network Simulation and Visualization Engine

V1.06

Table of Contents

DigiCortex Engine Introduction

DigiCortex is a highly optimized engine for large-scale simulation and interactive visualization of biologically plausible neural network circuits on heterogeneous computing architectures. The DigiCortex engine simulates (Figure 1) activity of neurons and synapses by using biologically accurate models of neurons, synapses and synaptic plasticity.

Highly optimized compute kernels enable simulations with millions of neurons simulated in real-time [1]. [2] DigiCortex engine has also been used to achieve the largest thalamocortical system simulation on a home PC, containing 16.7 million multi-compartment neurons with 4 billion synapses [3]. Thanks to its ability to highly utilize underlying hardware, the DigiCortex engine is also used by IT website AnandTech as one of CPU benchmarking tools [4].

This white paper provides a high level overview of DigiCortex features and gives an overview of achieved simulation performance results.

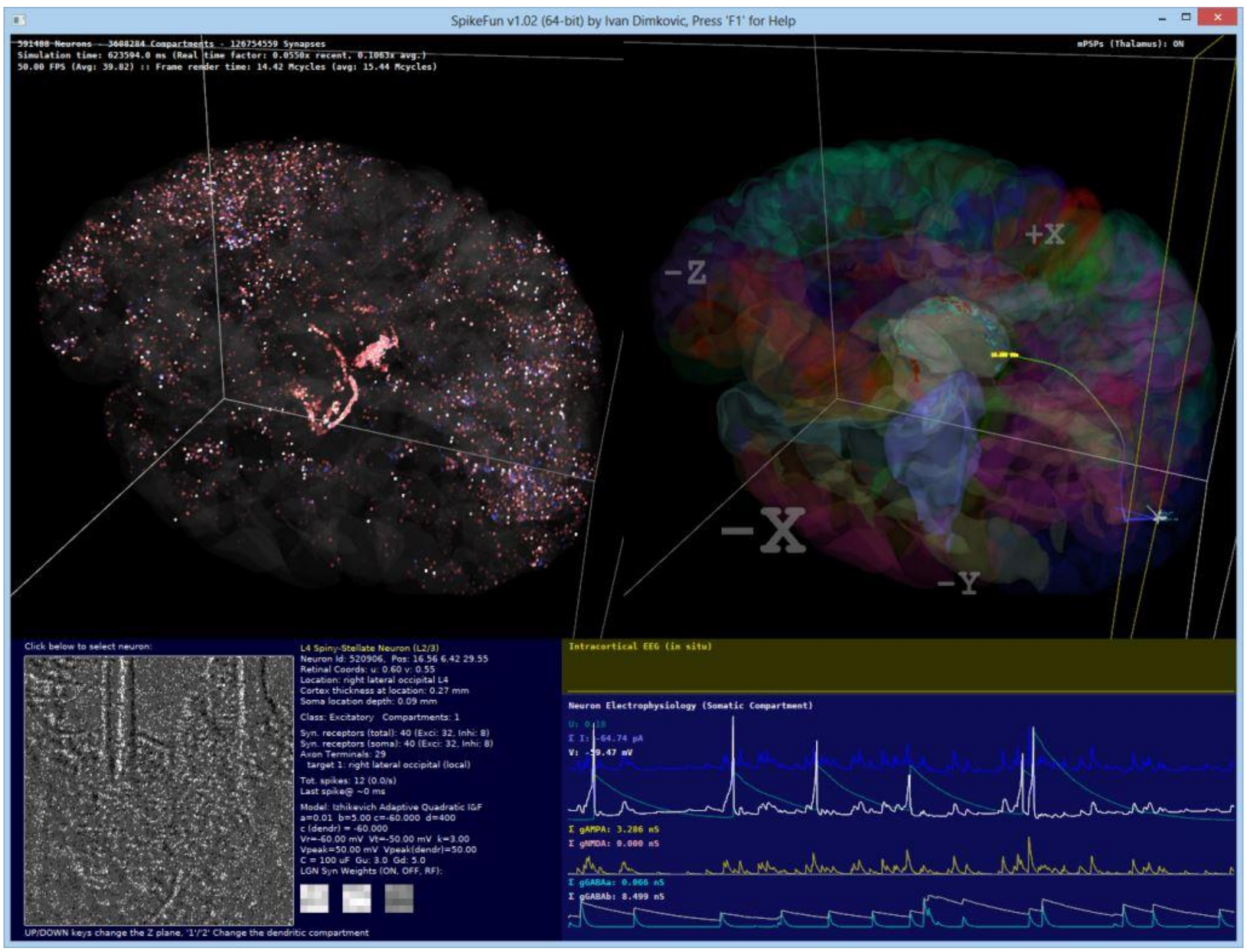

Figure 1 – DigiCortex Visualization Engine

Features

Table below provides a high-level summary of the DigiCortex engine features:

Supported Neuron Models	• **Adaptive Quadratic Integrate and Fire ("Izhikevich")* [5]** • **Adaptive Exponential Integrate and Fire ("AdEx")** [6]** *(*), (**) Neuron models are supported as single-compartment ("point") neurons as well as multiple-compartment neurons with arbitrary geometry, variable connectivity and membrane parameters per compartment. Other neuron models (e.g. Hodgkin-Huxley, Integrate and Fire, etc.) are possible as a future extension.*
Supported Neuron Types	**Major families:** • **Cortical:** Pyramidal, Spiny-Stellate, Basket and Non-Basket • **Thalamic:** Thalamocortical Relay, Thalamic Interneuron, Reticular Thalamic **Subfamilies:** • **Cortical:** Layer and Projection Specific (e.g. Pyramidal L2/3, Pyramidal L4, Pyramidal L5 (L2/3), Pyramidal L5 (L5/6), etc. • **Thalamic:** Specific and Non-Specific; Dorsal LGN ON/OFF Relay Cells (for simulations including retinal input) *Both major families and subfamilies are fully configurable and extensible (XML format) so new types of neurons can be defined. Each neuron sub-family has its own configurable membrane properties, geometry and connectivity statistics as well as myelinated and non-myelinated axonal spike propagation velocities.*
Supported Synaptic Receptor Types	• **Excitatory**: AMPA, NMDA • **Inhibitory**: GABAa, GABAb
Supported Synaptic Plasticity Models	• Short Term: Tsodyks-Markram phenomenological model [7]. Depressing, Facilitating and Pseudo-Linear Synapses are supported including differential signaling via the same axon [8] and can be fully defined and extended on per-simulation project basis. • Long Term: Voltage-based Spike-Timing-Dependent Plasticity (STDP / vSTDP) [9] with fully configurable parameters for each neuron family • Very Long Term: Synaptic metaplasticity [10] adapting synaptic weights in order to match desired neuron average spiking frequency
Visualization Features	• Interactive OpenGL rendering of live simulations • Real-time neuron picking and plotting of its connectivity and state • Simulated multi-site Intracranial EEG • Simulated fMRI BOLD • FFT analysis of spiking statistics and synaptic weight tracking
Circuit Builder	• Generation of arbitrary size thalamocortical neural circuits based on anatomical (T1/T2) and diffusion MRI data, statistical per-neuron family connectivity and cortical layer thickness

Table 1 – DigiCortex Features

An Overview of DigiCortex Architecture

DigiCortex engine is written in C++ language and designed to be modular and extensible (Figure 2). The actual simulation computation is performed in compute plug-ins, two of which are available today (Intel x86 and NVIDIA CUDA) and others can be implemented in future (e.g. FPGA or neuromorphic hardware).

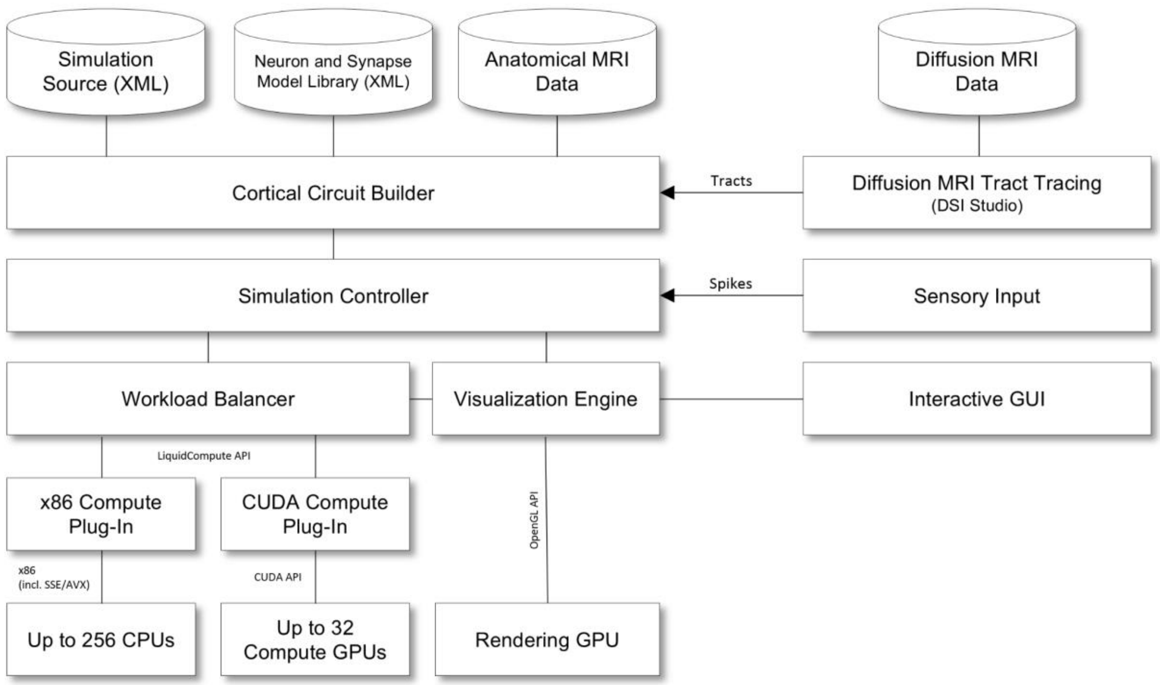

Figure 2 – DigiCortex Architecture

DigiCortex architecture consists of the following modules:

- **Cortical Circuit Builder** – Using anatomical (T1/T2) and diffusion (DTI/GQI/DSI) data as well as statistical connectivity rules from the simulation source project, cortical circuit builder creates the thalamocortical network. 3D placement of cortical and thalamic neurons is performed on the brain surface mesh generated using Freesurfer tool [11] based on anatomical T1/T2 MRI data. Local neuron connectivity is controlled by the simulation XML configuration and reference simulation projects use statistical connectivity data from cat's primary visual cortex [12]. Long-range neuron connectivity is generated using diffusion MRI data. Reference projects use high-resolution multi-shell diffusion MRI scans obtained by Human Connectome Project [13].

- **Simulation Controller** – Simulation controller is responsible for the lifecycle management of a simulation and also acts as an interface to an external spiking sensory input. The DigiCortex engine also implements a reference visual sensory input using Inria's Virtual Retina simulator [14]. Using the Virtual Retina module, it is possible to feed the simulations with live camera input and model the early visual system including ON/OFF retinal ganglion cells, ON/OFF dorsal lateral geniculate nucleus neurons as well as primary visual cortex. Combined with a voltage-based STDP it is possible to simulate development of orientation preferences in the primary visual cortex neurons.

- **Workload Balancer** – module actually responsible for executing the simulations and designed for support of Heterogeneous Compute Architectures (HCAs). Workload balancer is aware of the underlying hardware layout (e.g. NUMA nodes, available CPUs and GPUs) and optimizes the network memory layout and size of work packages in order to maximize performance on the target system. The actual computation on the target CPU/GPU is performed by compute plug-ins interfacing with work manager using LiquidCompute API.

- **Visualization Engine** – DigiCortex features a highly optimized real-time OpenGL rendering engine for biological neural network simulations. Variable level of rendering detail is configurable by user as a trade-off between rendering performance and level of details, supporting rendering of individual neuron compartment membrane parameters or spikes only. In addition, visualization engine implements simulated fMRI BOLD and multi-site Intracranial EEG.

- **Compute Plug-Ins** – in order to take advantage of evolving compute architectures and platforms, the DigiCortex engine features a plug-in interface allowing implementing support for different compute back-ends. This allows the actual computation to be done either on a local CPU, local GPU over the PCI Express (or equivalent bus) or, possibly, on a different type of local or remote accelerator. Following compute Plug-Ins are already supported in DigiCortex:

    - **x86 Compute** – supporting up to 256 Intel x86 CPUs and extensively hand optimized using SSE, AVX / AVX-2 and AVX-512 instructions. Selected parts of compute kernels were hand written in assembly code in order to extract maximum performance. x86 Compute Plug-In is NUMA aware and will maximize memory locality for the best performance.

    - **CUDA Compute** – supporting up to 32 NVIDIA GPUs (Kepler or later). Compute kernels were optimized for highly efficient execution of the GPU hardware and will take advantage of the high memory bandwidth available on modern NVIDIA GPUs.

Simulation Performance

Performance was one of the core design principles behind DigiCortex engine. Thanks to hand-optimized CPU and GPU code, DigiCortex engine is able to achieve extremely high simulation performance without sacrificing biological level of detail.

On a single NVIDIA Pascal GP100 accelerator (Tesla P100), with the current version of the DigiCortex engine (v1.22) it is possible to simulate 2.65 million single-compartment neurons with 105 million synapses in real-time including short-term and long-term synaptic plasticity as well as 1ms simulation time step. Using a multi-GPU platform, a real-time simulation performance of 12.6 million neurons and 431.6 million synapses was achieved on 8-way NVIDIA V100 ("Volta") system [15].

The table below indicates achieved real-time performance on different target platforms (please see Figure 3 for simulation configuration):

Platform Name	Num. CPUs/GPUs	Total Available Memory Bandwidth *(Theoretical)*	Max. Compute Performance *(Theoretical, FP32)*	Real-time Simulation Performance *(1 ms simulation time = 1 ms real-time)*
Dual Intel Xeon E5-2699 v3 ("Haswell EP")	2 (36 CPU cores)	136 GB/s	1.9 TFLOPS[1]	**0.21M Neurons 8.3M Synapses**
Single NVIDIA GP102 ("Pascal")	1 (3584 CUDA cores)	484 GB/s	11.3 TFLOPS	**1.8M Neurons 70M Synapses**
Single NVIDIA P100 ("Pascal")	1 (3584 CUDA cores)	720 GB/s	10.6 TFLOPS	**2.6M Neurons 105M Synapses**
Dual NVIDIA GP100 ("Pascal")	2 (7168 CUDA cores)	1440 GB/s	21.2 TFLOPS	**5.0M Neurons 201M Synapses**
8-Way NVIDIA GK210 ("Kepler")	16 (39936 CUDA cores)	3840 GB/s	69.9 TFLOPS	**4.4M Neurons 153.7M Synapses**
8-Way NVIDIA V100 ("Volta")	8 (43008 CUDA cores)	7200 GB/s	120 TFLOPS	**12.6M Neurons 431.6M Synapses**

Table 2 – Number of neurons / synapses simulated in real-time on different platforms

[1] Calculated as: CPU AVX Base Frequency * Number of Cores * 16 FLOPS per cycle

Chart below shows simulation performance scaling across different simulation sizes and on two generations of GPU platform:

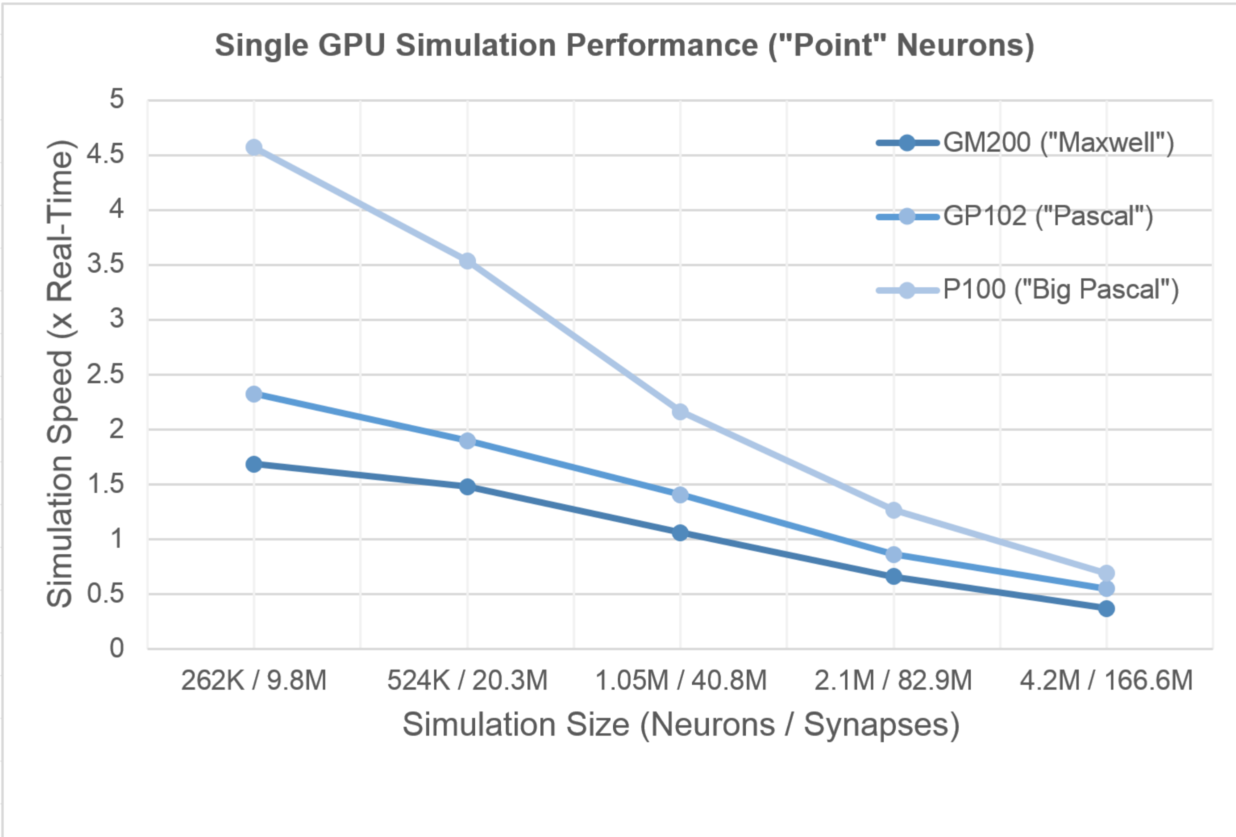

Figure 3 – DigiCortex Simulation Performance Scaling on GPU platform

Integration Step Size	1 ms
Numerical Solver	Forward + Backward Euler Method
Neuron Models	Single and multiple compartment Adaptive Quadratic Integrate & Fire ("Izhikevich" [5])
Neuron Types Simulated	Excitatory: Pyramidal, Spiny Stellate, Thalamocortical Relay; Inhibitory: Basket, Non-basket, Reticular Thalamic, Thalamic Interneuron
Number of cells	Test case specific, please refer to Table 2
Excitatory / Inhibitory Cell Ratio	77% Excitatory, 23% Inhibitory
Synaptic Receptor Types	AMPA, NMDA, GABAa, GABAb
Short-Term Synaptic Plasticity	Yes (Tsodyks & Markram model [7]), enabled for all synapses
Long-Term Synaptic Plasticity	Yes (Voltage-based STDP [9]), enabled for excitatory synapses
Network Geometry	Thalamocortical system with 6-layer cortex, 3D cell positioning based on anatomical MRI data, axonal guidance using Diffusion MRI data

Figure 4 – Benchmark Simulation Configuration

Conclusion and Future Work

Large-scale simulations of biological neural networks are becoming the key to understanding brain computation and help solve numerous diseases involving nervous system, as evidenced by the large scientific initiatives in Europe (Human Brain Project [16]) and USA (BRAIN initiative [17]).

DigiCortex Engine is a highly scalable simulation and visualization engine for biological neural networks allowing "digital experiments" on synthetic neural circuits on a very large scale (10^7 neurons, 10^9 synapses on a single compute node) previously impossible outside of supercomputing resources. Emergence of the readily available massive cloud computing resources is one of the area that could enable scaling of simulations to the numbers comparable to mammalian brains. This is one of the area Accelerico team is exploring.

Contact us (contact@accelerico.com) for more details and inquiries about the DigiCortex engine.

References

[1] Accelerico, "DigiCortex 1.21 - Support for NVIDIA Pascal - 1.8 Million Neurons Simulated in Real-Time," 20 April 2017. [Online]. Available: http://www.digicortex.net/node/51.

[2] I. Dimkovic, "DigiCortex NVIDIA P100 ("Pascal") Update: 5M Neurons Simulated in Real-time on Dual GP100 GPU," 06 June 2017. [Online]. Available: http://www.digicortex.net/node/53.

[3] I. Dimkovic, "4 Billion Synapses, 16.7 Million Neurons, Visual System," 23 April 2014. [Online]. Available: http://www.digicortex.net/node/47.

[4] AnandTech, "2017 CPU Tests (Win10)," 2017. [Online]. Available: http://www.anandtech.com/bench/CPU/1616.

[5] E. M. Izhikevich, "Simple Model of Spiking Neurons," *IEEE Transactions on Neural Networks,* vol. 14, no. 6, pp. 1569-1572, November 2003.

[6] R. Brette and W. Gerstner, "Adaptive Exponential Integrate-and-Fire Model as an Effective Description of Neuronal Activity," *Journal of Neurophysiology,* vol. 94, pp. 3637-3642, 1 November 2005.

[7] M. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proceedings of the National Academy of Sciences,* pp. 719-723, February 1997.

[8] H. Markram, Y. Wang and T. M., "Differential signaling via the same axon of neocortical pyramidal neurons," *Proc Natl Acad Sci U S A,* pp. 5323-5328, 28 April 1998.

[9] C. Clopath, L. Büsing, E. Vasilaki and W. Gerstner, "Connectivity reflects coding: a model of voltage-based STDP with homeostasis," *Nature Neuroscience,* pp. 344-352, 24 January 2010.

[10] K. Carlson, M. Richert, N. Dutt and J. L. Krichmar, "Biologically plausible models of homeostasis and STDP: Stability and learning in spiking neural networks," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013.

[11] A. Dale, B. Fischl and M. Sereno, "Cortical surface-based analysis. I. Segmentation and surface reconstruction," *Neuroimage,* vol. 9, pp. 179-194, February 1999.

[12] T. Binzegger, R. J. Douglas and K. A. Martin, "A Quantitative Map of the Circuit of Cat Primary Visual Cortex," *Journal of Neuroscience,* vol. 39, pp. 8441-8453, 4 September 2004.

[13] K. Setsompop, R. Kimmlingen, E. Eberlein, T. Witzel, J. Cohen-Adad, J. McNab, B. Keil, M. Tisdall, P. Hoecht, P. Dietz, S. Cauley, V. Tountcheva, V. Matschl, V. Lenz, K. Heberlein, A. Potthast, H. Thein, J. Van Horn, A. Toga, F. Schmitt, D. Lehne, B. Rosen, V. Wedeen and L. Wald, "Pushing the limits of in vivo diffusion MRI for the Human Connectome Project," *NeuroImage,* vol. 80, pp. 220-233, 15 October 2013.

[14] A. Wohrer and P. Kornprobst, "Virtual Retina: A biological retina model and simulator, with contrast gain control," vol. 26, no. 2, pp. 219-249, April 2009.

[15] I. Dimkovic, "DigiCortex NVIDIA V100 ("Volta") Update: 12.7M Neurons Simulated in Real-time on 8x V100 GPU," 10 October 2017. [Online]. Available: http://www.digicortex.net/node/54. [Accessed 10 October 2017].

[16] H. Markram, K. Meier, T. Lippert, S. Grillner, R. Frackowiak, S. Dehaene, A. Knoll, H. Sompolinsky, K. Verstreken, J. DeFelipe, S. Grant, J. Changeux and A. Saria, "Introducing the Human Brain Project," in *Procedia Computer Science*, 2011.

[17] White House Office of the Press Secretary, "Fact Sheet: BRAIN Initiative," 2 04 2013. [Online]. Available: https://obamawhitehouse.archives.gov/the-press-office/2013/04/02/fact-sheet-brain-initiative.

Notice

ALL INFORMATION PROVIDED IN THIS WHITE PAPER, INCLUDING COMMENTARY, OPINION, DRAWINGS, LISTS AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED "AS IS". ACCELERICO MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied.

Trademarks

Accelerico, Accelerico Logo, DigiCortex and LiquidCompute are trademarks or registered trademarks of Accelerico. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright