Tutorial and Live Demo

Accelerating HPC and AI Applications using MVAPICH2-DPU, X-ScaleHPL-DPU, and X-ScaleAI-DPU Packages

Donglai Dai and Kyle Schaefer

X-ScaleSolutions

http://x-scalesolutions.com

Overview of X-ScaleSolutions

- Started in 2018, bring innovative and efficient end-to-end solutions, services, support, and training to our customers
- Commercial support and training for the state-of-the-art communication libraries
  - Platform-specific optimizations and tuning
  - Application-specific optimizations and tuning
  - Obtaining guidelines on best practices
  - Timely support for installation and operational issues encountered with the library
  - Flexible Service Level Agreements
  - Web portal interface to submit issues and tracking their progress
  - Information on major releases and periodic information on major fixes and updates
  - Help with upgrading to the latest release
- Winner of multiple U.S. DOE SBIR grants
- Market these products for HPC and AI applications with commercial support
- A Silver ISV member of the OpenPOWER Consortium

Overview of X-ScaleSolutions (cont'd)

- Currently, we offer five products with commercial support:
  - MVAPICH2-DPU (https://x-scalesolutions.com/mvapich2-dpu/)
  - X-ScaleHPC (https://x-scalesolutions.com/x-scalehpc/)
  - X-ScaleHPL-DPU (https://x-scalesolutions.com/xscale-hpl-dpu/)
  - X-ScaleAI (https://x-scalesolutions.com/x-scaleai/)
  - X-ScaleAI-DPU (https://x-scalesolutions.com/x-scaleai-dpu/)
- More information about the specific features and capabilities of these products are available on the websites provided above
- Today's demo will focus on the three DPU related products

X-ScaleSolutions will give a presentation at 3 pm ET on Wednesday Aug 23 that will go into more performance results of our products. Please come and join us.

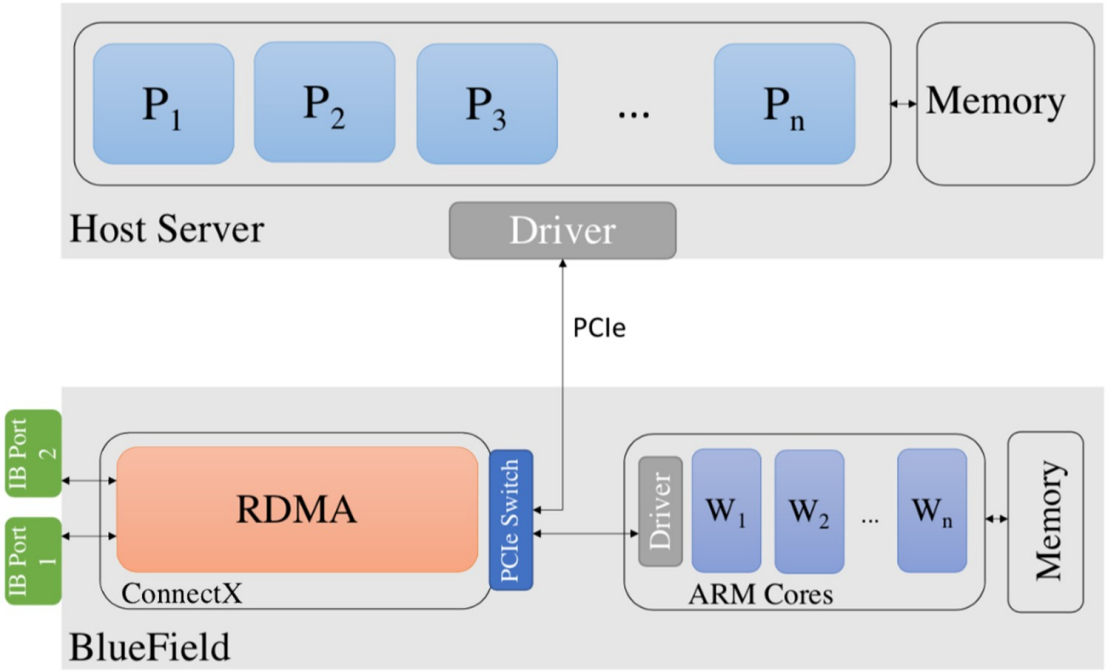
Requirements for Next-Generation MPI Libraries

- Message Passing Interface (MPI) libraries are used for HPC and AI applications
- Requirements for a high-performance and scalable MPI library:
  - Low latency communication
  - High bandwidth communication
  - Minimum contention for host CPU resources to progress non-blocking collectives
  - High overlap of computation with communication
- CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation

Network offload mechanisms are gaining attraction as they have the potential to completely offload the communication of MPI primitives into the network

Overview of BlueField-2/3 DPU

- ConnectX-6 network adapter with 100Gbps/200Gbps InfiniBand
- System-on-chip containing 8/16 64-bit ARMv8 A72/A76 cores with 2.75 GHz each
- 16/32 GB of memory for the ARM cores



How to Re-design an MPI library to take advantage of DPUs and accelerate scientific applications?

MVAPICH2-DPU Library 2023.05 Release

- Released in May 2023
- Based on MVAPICH2 2.3.7
- Supports all features available with the MVAPICH2 2.3.7 release (http://mvapich.cse.ohio-state.edu)
- Novel framework to offload non-blocking collectives to DPU
- Supports offloads of the following non-blocking collectives
  - Alltoall (MPI_Ialltoall)
  - Broadcast (MPI_Ibcast)

MVAPICH2-DPU Library 2023.05 Release (Cont'd)

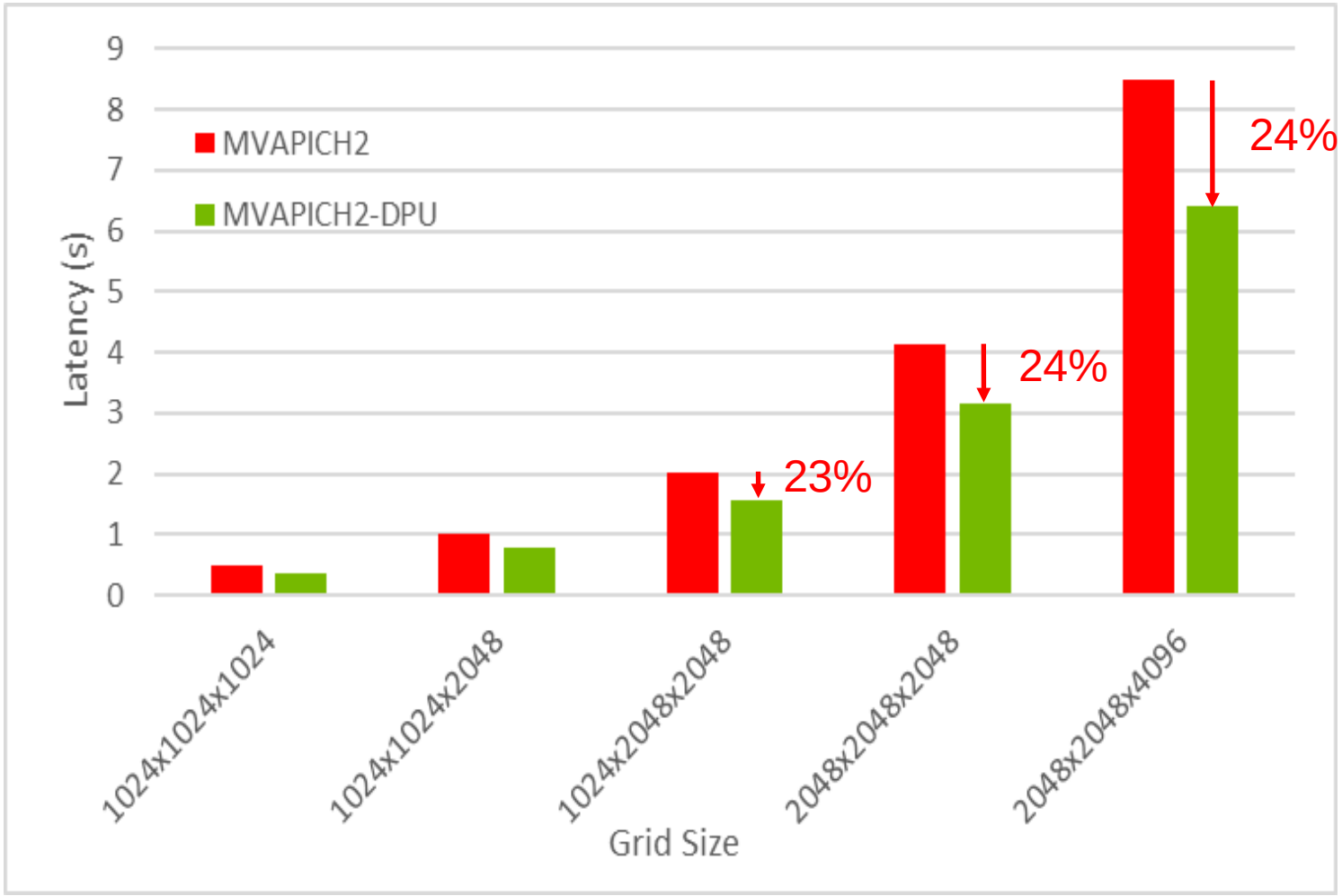
- Significantly increases (up to 100%) overlap of computation with any mix of MPI_Ialltoall or MPI_Ibcast non-blocking collectives
- Accelerates scientific applications using any mix of MPI_Ialltoall or MPI_Ibcast non-blocking collectives

Available from X-ScaleSolutions, please send a note to contactus@x-scalesolutions.com to get a trial license.

P3DFFT Application Execution Time (32 nodes), BF-2 100Gbps, Intel Platform

32 nodes with 32 ppn
(1,024 processes)

32x32 process grid

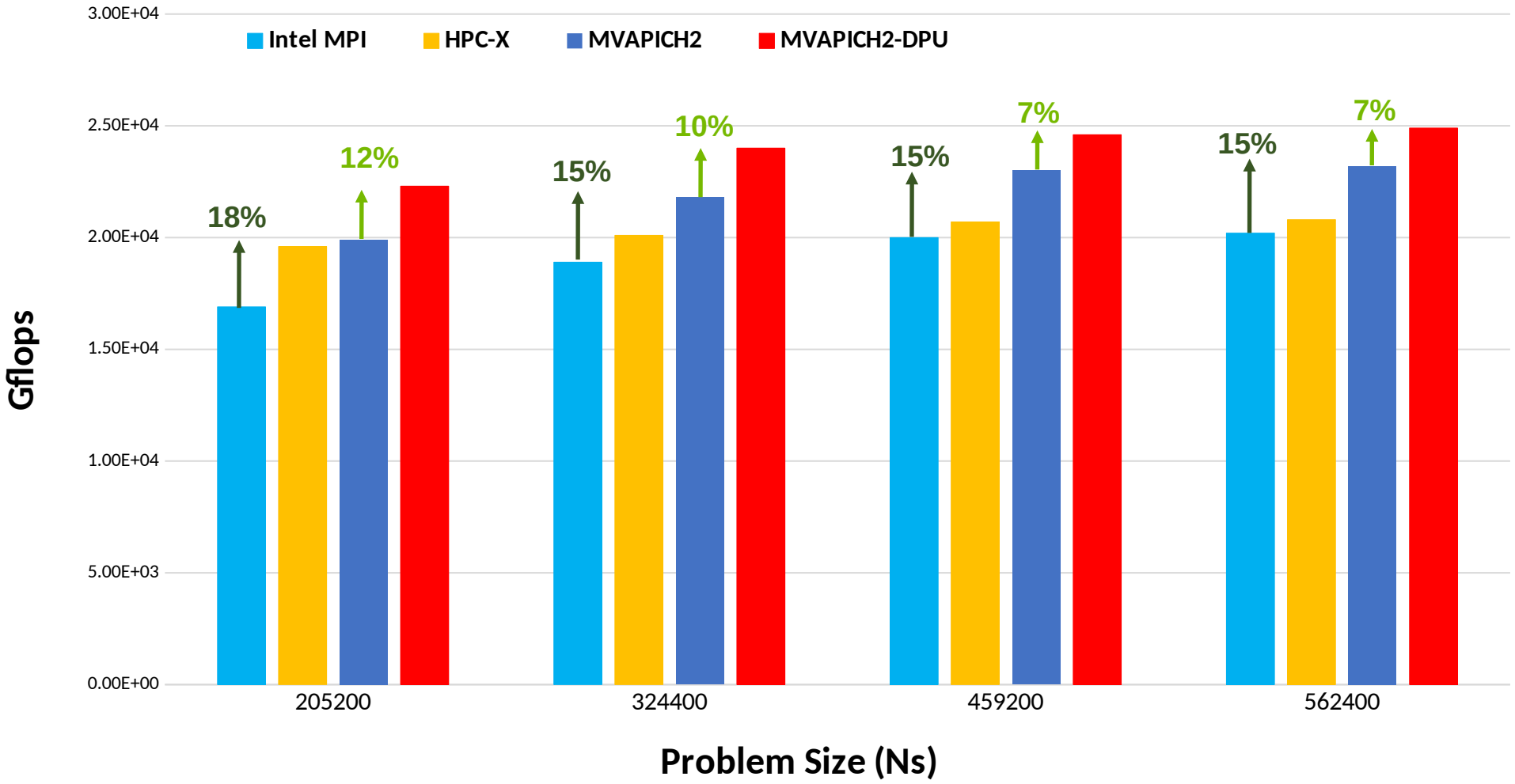


Benefits in application-level execution time

X-ScaleHPL-DPU Package 2023.05 Release

- Released in May 2023
- Based on High-Performance Linpack Code (HPL) v2.3
- Codesigned with MVAPICH2-DPU v2023.05 library
  - Supports two modes: DPU mode and Host mode
  - In DPU mode: the benchmark application intelligently offloads non-blocking broadcast (MPI_Ibcast) operations to DPU
  - In Host mode: no such offloading occurs

HPL Benchmark Performance (8 EPYC nodes, 128 ppn)



Performance benefits at application-level

X-ScaleAI-DPU Package 2023.05 Release

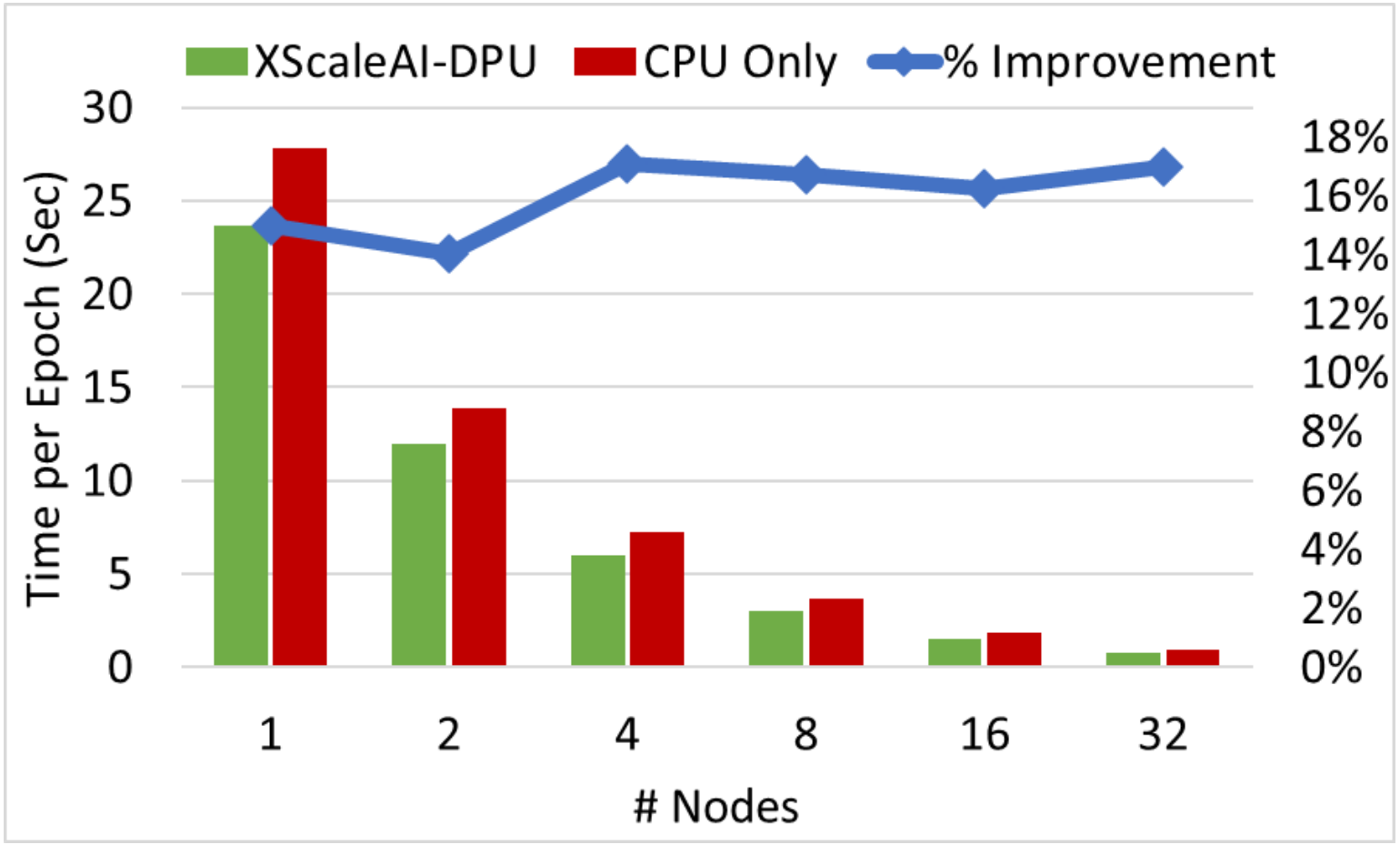
High performance solution to accelerate CPU-based Deep Learning training by utilizing the capabilities of DPUs

- Released in May 2023
- Distributed Training with PyTorch using Horovod, based on PyTorch v1.12.0 and Horovod v0.25.0
- Co-designed with MVAPICH2-DPU library 2023.05 release
- Offload DNN training tasks to the DPU
- User friendly Python interface to run DL applications, simple installation and execution using one command for each
- “Out of the box” optimal performance on CPU+DPU platforms
- Tested using popular DNN models and datasets with up to 17% improvement in performance

Training of ResNet-20v1 model on CIFAR10 dataset

System Configuration

- Two Intel(R) Xeon(R) 16-core CPUs (32 total) E5-2697A V4 @ 2.60 GHz
- NVIDIA BlueField-2 SoC, HDR100 100Gb/s InfiniBand/VPI adapters
- Memory: 256GB DDR4 2400MHz RDIMMs per node
- 1TB 7.2K RPM SSD 2.5" hard drive per node
- NVIDIA ConnectX-6 HDR/HDR100 200/100Gb/s InfiniBand/VPI adapters with Socket Direct



Performance improvement using X-ScaleAI-DPU over CPU-only training on the ResNet-20v1 model on the CIFAR10 dataset

Today's Live Demo

- Being run on the HPC-AI Advisory Council cluster
  - 32 Xeon nodes connected with 32 DPUs over 200Gbps InfiniBand
  - 1,024 CPU cores (Xeons) and 256 ARM cores (DPUs)
- Configuration
  - Server HW:
    - CPU: Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz
    - Adapter: Nvidia BlueField-2 DPU, 8 ARM cores 2.75 Ghz, 16GB DDR4
  - Software/Firmware:
    - OS version: CentOS 8.3
    - Driver version: 5.2-1
    - Firmware version : 24.30.1004
  - MPI:
    - MVAPICH2-DPU 2023.05
  - OSU Micro-Benchmarks (OMB) 5.7.1
  - P3DFFT application v2.3

Today's Live Demo (Cont'd)

- Five parts on performance benefits
  - OSU MPI Micro-Benchmarks (OMB 5.7.1) with Ialltoall
  - P3DFFT application (using non-blocking Alltoall)
  - OMB with Ibcast
  - X-ScaleHPL-DPU release 2023.05 (using non-blocking Broadcast)
  - X-ScaleAI-DPU release 2023.05

Future Releases and Engagement Plan

- Upcoming Support to Non-blocking Alltoallv Using DPU
  - Up to 60% performance improvement in OMB Ialltoallv benchmark tests with DPU offloading vs without (i.e., host only)
- Offloading designs for other non-blocking collectives
  - Allreduce, Reduce, etc.
- Offloading designs for other MPI functions
- Application-level and scalability studies
- Co-designing MPI and AI applications with DPU support

X-ScaleSolutions will be happy to get engaged, please send a note to contactus@x-scalesolutions.com.

Thank You!

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/