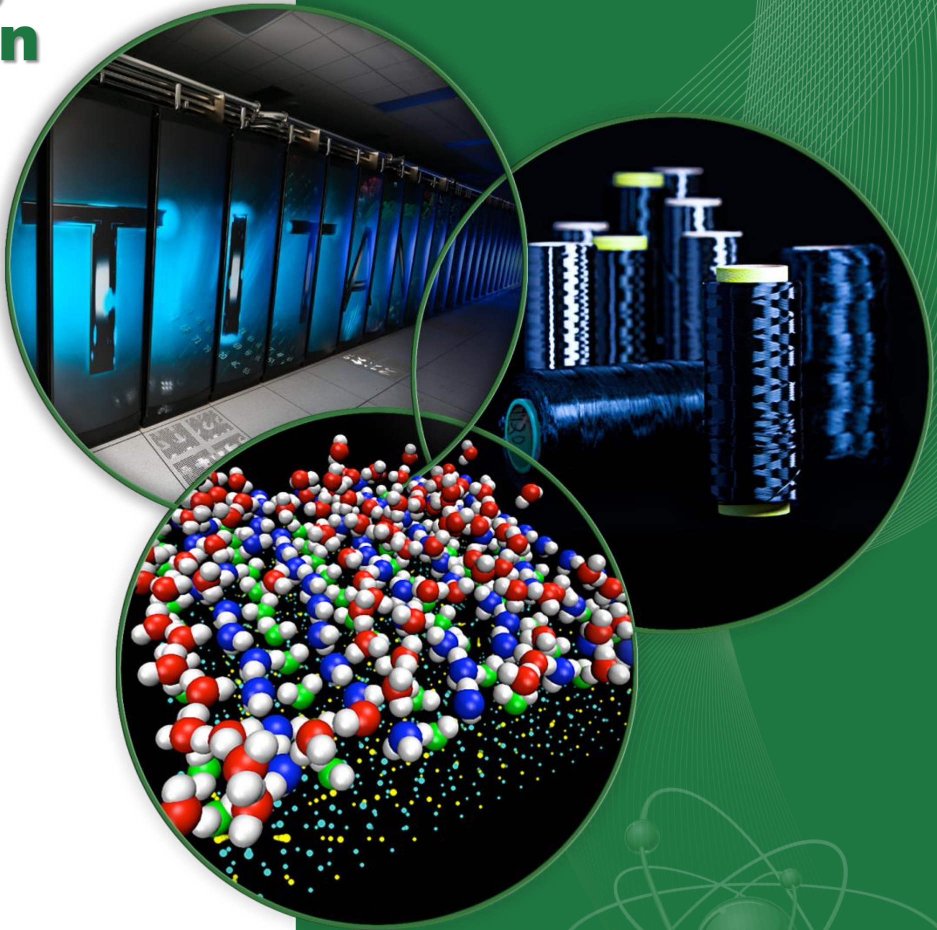
Dense/sparse numeric tensor algebra: Scalable, hardware-agnostic design for performance portability



Dmitry I. Lyakh

liakhdi@ornl.gov

Tensors

- Vector space ***A***:

$$\boldsymbol{A} \equiv span(\{a_i | i = 1 \dots m\})$$

- Vector space ***B***:

$$\boldsymbol{B} \equiv span(\{b_j | j = 1 \dots n\})$$

- Direct-product (tensor-product) space ***C***:

$$\boldsymbol{C} \equiv span\left(\left\{c_{ij} \equiv a_i b_j \middle| i = 1 \dots m, j = 1 \dots n\right\}\right)$$

- Dual basis:

$$a^i = g^{ij} a_j \qquad \overset{\text{Inner product}}{g_{ij} \equiv (a_i, a_j)} \qquad g^{ij} = (g_{ij})^{-1}$$

- Any vector space with a direct-product structure is represented by tensors

Applications

- Quantum many-body theory: Electronic and nuclear structure, quantum computing: Wave-function
- Loop quantum gravity: Fine structure of space-time
- Multivariate data analysis: Extracting correlations, features, data compression
- Classical physics
- Any data arranged as a multi-dimensional array is a tensor (in the most general sense)
- Any function of multiple variables discretized over a grid or some functional inner-product space is a tensor

CAAR Program @OLCF: Chemistry

- **DIRAC** (PI L. Visscher): Relativistic coupled-cluster theory
- **LS-DALTON** (PI P. Jorgensen): Local non-relativistic coupled-cluster theory for large molecular systems
- **NWChem** (PI K. Kowalski): Non-local non-relativistic coupled-cluster theory
- <u>Non-CAAR</u>: **ACES-IV** (PI R. Bartlett): Non-local non-relativistic coupled-cluster theory
- Other quantum-chemistry codes…

MATH: DENSE/BLOCK-SPARSE TENSOR ALGEBRA

Coupled-Cluster Theory

$$|\Psi\rangle = \exp(\hat{T})|0\rangle = \left(1 + \hat{T} + \frac{1}{2!}\hat{T}^2 + \frac{1}{3!}\hat{T}^3 + \frac{1}{4!}\hat{T}^4 + \cdots\right)|0\rangle$$

$$|\Psi_{excited}\rangle = \hat{R}e^{\hat{T}}|0\rangle$$

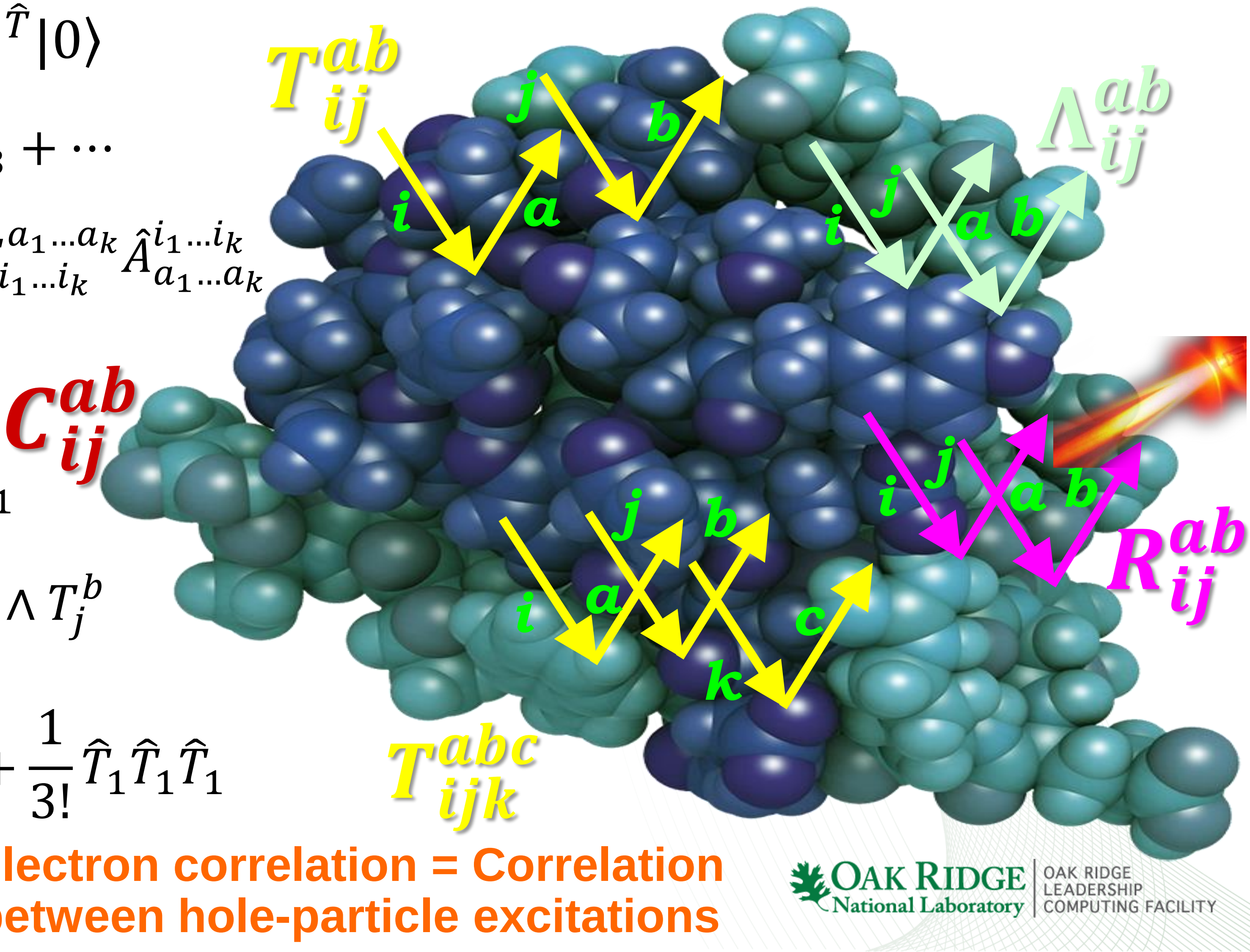
$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots$$

$$\hat{T}_k = \frac{1}{k!\,k!}\sum_{\substack{a_1\ldots a_k\\ i_1\ldots i_k}} T^{a_1\ldots a_k}_{i_1\ldots i_k}\hat{A}^{i_1\ldots i_k}_{a_1\ldots a_k}$$

$$\hat{C}_2 = \hat{T}_2 + \frac{1}{2!}\hat{T}_1\hat{T}_1$$

$$C^{ab}_{ij} = T^{ab}_{ij} + T^a_i \wedge T^b_j$$

$$\hat{C}_3 = \hat{T}_3 + \hat{T}_2\hat{T}_1 + \frac{1}{3!}\hat{T}_1\hat{T}_1\hat{T}_1$$



Electron correlation = Correlation between hole-particle excitations

Math Framework: Basic Tensor Algebra

- Formal tensor: $T^{pq...}_{rs...} \mapsto T(p,q,r,s,...)$: n-D Array

  Full tensor: $T(p,q,r,s): \; p \in P, q \in Q, r \in R, s \in S$

  Tensor slice: $T(p,q,r,s): \; p \in P' \subseteq P, q \in Q' \subseteq Q,$

  $r \in R' \subseteq R, s \in S' \subseteq S$

Few primitive operations:

- Tensor addition:

$$\forall p,q,r,s: \;\; T^{pq}_{rs} = L^{pq}_{rs} + R^{pq}_{rs}$$

Parallelism!

- Tensor product:

$$\forall p,q,r,s: \;\; T^{pq}_{rs} = L^{p}_{r} R^{q}_{s}$$

Parallelism!

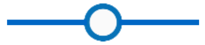
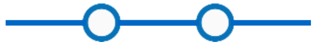
- Tensor contraction:

$$\forall p,q,r,s: \;\; T^{pq}_{rs} = L^{pai}_{bcd} R^{qbcd}_{rsai}$$

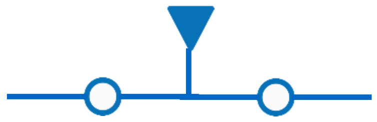
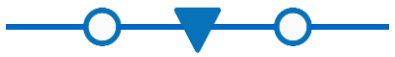
Compute intensive (potentially)!

Math Framework: Tensor Decompositions

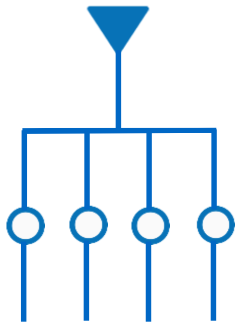
- Graphical (diagrammatic) representation:

Matrix:  Matrix*Matrix: 

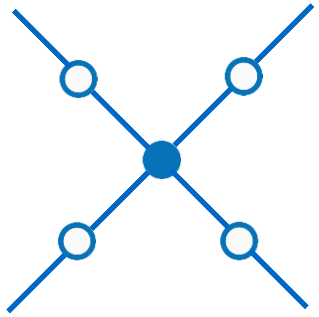
- Linear algebra: SVD is optimal in the 2-norm:

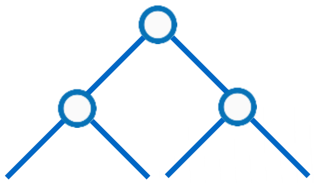
- Tensor (multi-linear) algebra: Many choices:



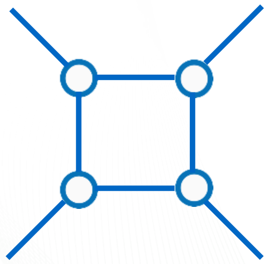
Canonical polyadic



Tucker



Tensor tree



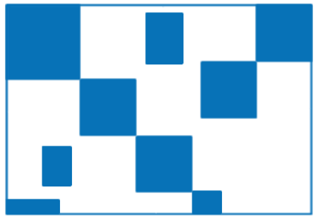
Tensor train

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

Tensor Sparsity

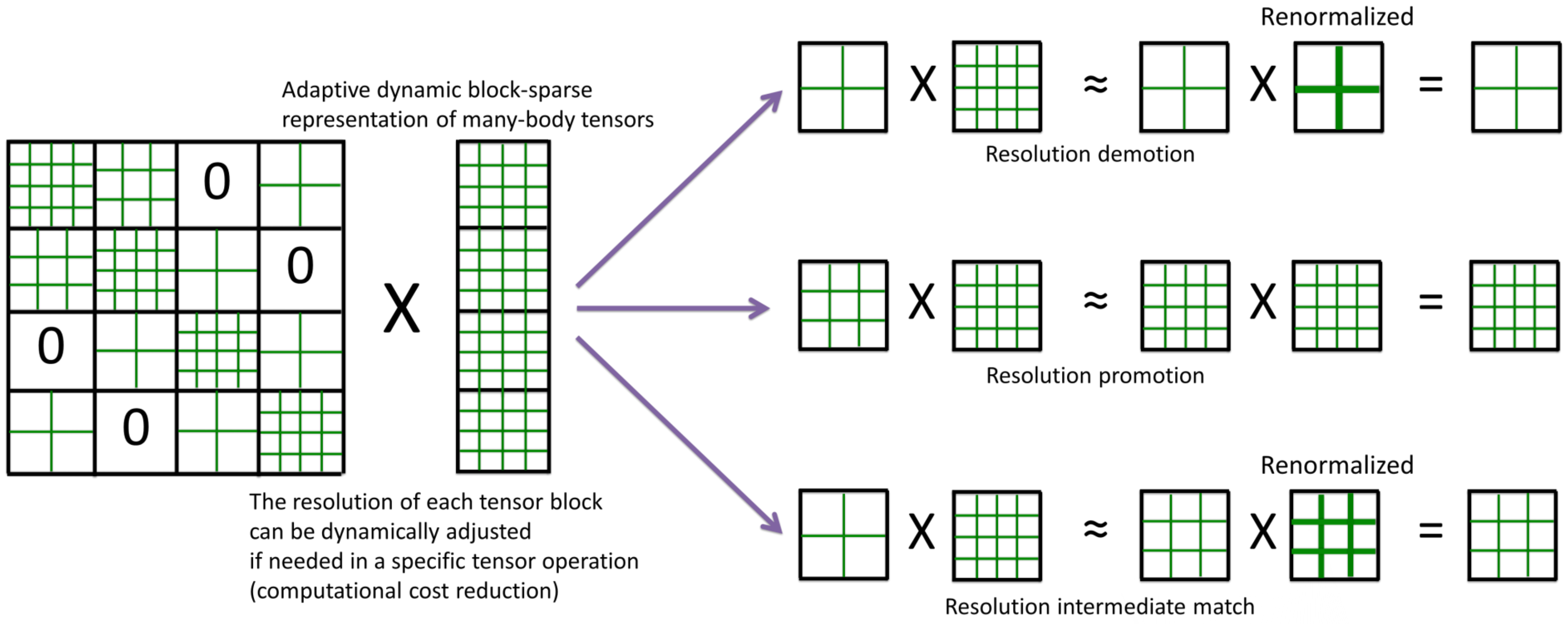
- Dense tensors:

- Block-sparse tensors:

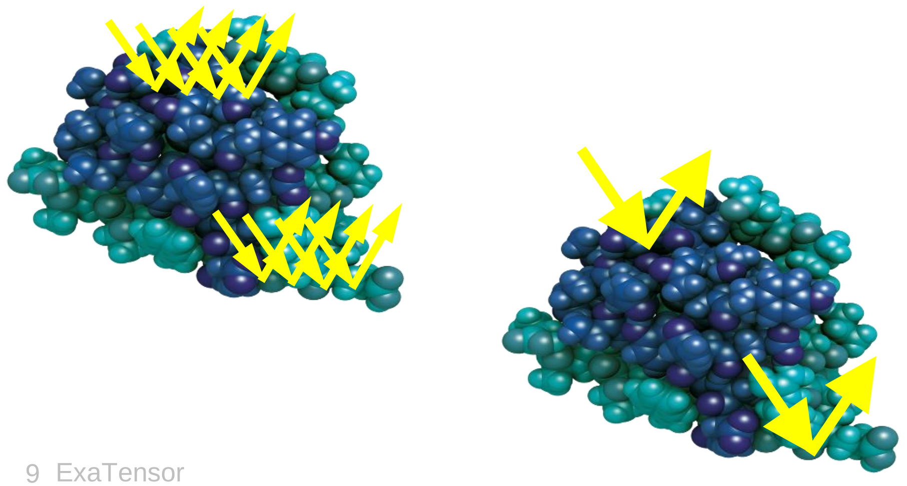


- Regular sparse tensors:
  - Some regular sparsity pattern other than block sparsity
- Irregular sparse tensors:
  - Sparsity pattern has no regularity

Resolution-Adaptive Tensor Algebra


Adaptive dynamic block-sparse representation of many-body tensors
0
0
0
0
X
Renormalized
X
≈
X
=
Resolution demotion
X
≈
X
=
Resolution promotion
Renormalized
X
≈
X
=
Resolution intermediate match

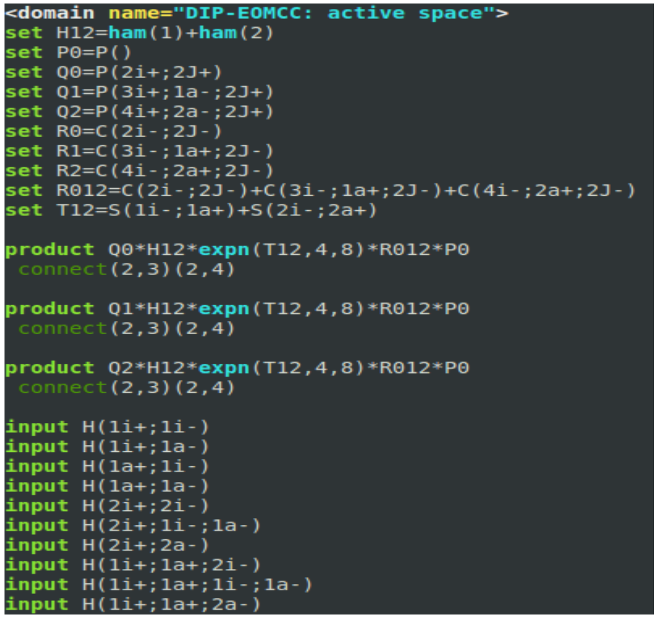
The resolution of each tensor block can be dynamically adjusted if needed in a specific tensor operation (computational cost reduction)



- Large tensor elements can become tensors themselves (higher resolution);
- Weak tensor slices can be compressed by lowering the resolution, up to a single (complex) number;
- Adapt to the calculated electronic state and available HPC resources;
- Should be better than just black-and-white discarding.

Automated Development Tools and DSL

- Automated equation/code generators are unavoidable in CC
- DSLangs and DSLibs can provide a scalable development environment: **DiaGen** input/output:



```
<domain name="DIP-EOMCC: active space">
set H12=ham(1)+ham(2)
set P0=P()
set Q0=P(2i+;2J+)
set Q1=P(3i+;1a-;2J+)
set Q2=P(4i+;2a-;2J+)
set R0=C(2i-;2J-)
set R1=C(3i-;1a+;2J-)
set R2=C(4i-;2a+;2J-)
set R012=C(2i-;2J-)+C(3i-;1a+;2J-)+C(4i-;2a+;2J-)
set T12=S(1i-;1a+)+S(2i-;2a+)

product Q0*H12*expn(T12,4,8)*R012*P0
 connect(2,3)(2,4)

product Q1*H12*expn(T12,4,8)*R012*P0
 connect(2,3)(2,4)

product Q2*H12*expn(T12,4,8)*R012*P0
 connect(2,3)(2,4)

input H(1i+;1i-)
input H(1i+;1a-)
input H(1a+;1i-)
input H(1a+;1a-)
input H(2i+;2i-)
input H(2i+;1i-;1a-)
input H(2i+;2a-)
input H(1i+;1a+;2i-)
input H(1i+;1a+;1i-;1a-)
input H(1i+;1a+;2a-)
```

(285) $192.3.896 : Z^{A^b_1}_{I^a_1 I^a_2 I^b_1} += H^{l^a_1 K^a_1}_{d^a_1, d^a_2} S^{d^a_1}_{I^a_1} S^{d^a_2}_{I^a_2} C^{A^b_1}_{I^b_1, l^a_1 K^a_1} \cdot +1/2$

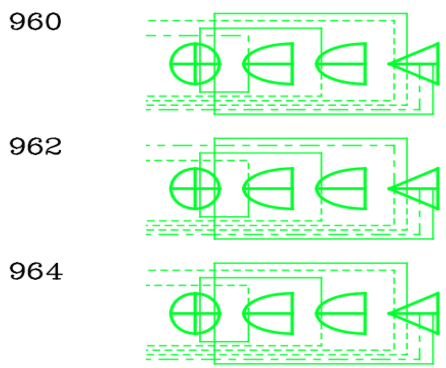
(286) $198.1.932 : Z^{A^b_1}_{I^a_1 I^a_2 I^b_1} += H^{l^b_1, l^b_2}_{d^b_1, d^b_2} S^{d^b_1}_{I^b_1} S^{d^b_2}_{l^b_1} C^{A^b_1}_{I^a_1 I^a_2, l^b_2}$

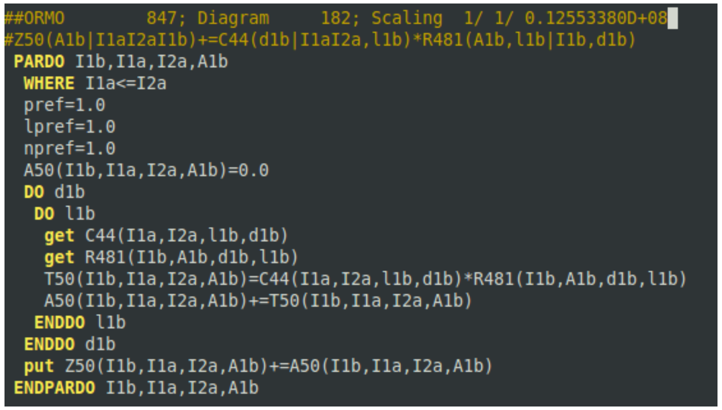
(287) $198.2.933 : Z^{A^b_1}_{I^a_1 I^a_2 I^b_1} += H^{l^b_1, l^a_1}_{d^a_1, d^b_1} S^{d^a_1}_{I^a_1} S^{d^b_1}_{l^b_1} C^{A^b_1}_{I^a_2 I^b_1, l^a_1}$

(288) $198.4.935 : Z^{A^b_1}_{I^a_1 I^a_2 I^b_1} += H^{l^a_1, l^b_1}_{d^b_1, d^a_1} S^{d^b_1}_{I^b_1} S^{d^a_1}_{l^a_1} C^{A^b_1}_{I^a_1 I^a_2, l^b_1}$

(289) $198.5.936 : Z^{A^b_1}_{I^a_1 I^a_2 I^b_1} += H^{l^a_1, l^a_2}_{d^a_1, d^a_2} S^{d^a_1}_{I^a_1} S^{d^a_2}_{l^a_1} C^{A^b_1}_{I^a_2 I^b_1, l^a_2}$

(290) $202.1.946 : Z^{A^b_1}_{I^a_1 I^a_2 I^b_1} += H^{l^b_1, K^a_1}_{d^a_1 d^b_1} S^{A^b_1}_{l^b_1} S^{d^a_1 d^b_1}_{I^a_1 I^b_1} C_{I^a_2, K^a_1}$





```
##ORMO        847; Diagram     182; Scaling  1/ 1/ 0.12553380D+08
#Z50(A1b|I1aI2aI1b)+=C44(d1b|I1aI2a,l1b)*R481(A1b,l1b|I1b,d1b)
 PARDO I1b,I1a,I2a,A1b
  WHERE I1a<=I2a
  pref=1.0
  lpref=1.0
  npref=1.0
  A50(I1b,I1a,I2a,A1b)=0.0
  DO d1b
   DO l1b
    get C44(I1a,I2a,l1b,d1b)
    get R481(I1b,A1b,d1b,l1b)
    T50(I1b,I1a,I2a,A1b)=C44(I1a,I2a,l1b,d1b)*R481(I1b,A1b,d1b,l1b)
    A50(I1b,I1a,I2a,A1b)+=T50(I1b,I1a,I2a,A1b)
   ENDDO l1b
  ENDDO d1b
  put Z50(I1b,I1a,I2a,A1b)+=A50(I1b,I1a,I2a,A1b)
 ENDPARDO I1b,I1a,I2a,A1b
```

(447) $324.85.1.1.3.1.0.20333376.09 : Z^{l^b_1}_{I^a_1 I^a_2 i^b_1} += H^{l^b_1, l^b_2}_{i^b_1, d^b_1} C^{d^b_1}_{I^a_1 I^a_2, l^b_2} \cdot -1.$

(448) $331.86.1.1.2.1.0.10042704.09 : Z^{l^b_1}_{I^a_1 I^a_2 i^b_1} += H^{l^b_1, K^a_1}_{I^a_1, d^b_1} C^{d^b_1}_{I^a_2 i^b_1, K^a_1} \cdot -1.$

(449) $325.85.1.1.3.1.0.20333376.09 : Z^{l^b_1}_{I^a_1 I^a_2 i^b_1} += H^{l^b_1, l^a_1}_{i^b_1, d^a_1} C^{d^a_1}_{I^a_1 I^a_2, l^a_1} \cdot -1.$

(450) $821.177.2.1.2.1.0.49593600.07 : Z^{l^b_1}_{I^a_1 I^a_2 i^b_1} += S^{d^b_1}_{i^b_1} R^{l^b_1}_{I^a_1 I^a_2, d^b_1} \cdot -1.$

(451) $938.199.1.2.2.2.0.11716488.10 : R^{l^b_1, K^a_1}_{I^a_1 i^b_1} += H^{l^b_1, K^a_1}_{d^a_1 d^b_1} S^{d^a_1 d^b_1}_{I^a_1 i^b_1}$

Efficient flexible runtime?

Computational Challenges

- Dense tensor algebra:
  - Communication bandwidth: Communication avoiding and regularization (similar to the matrix multiplication)
  - Memory size: Out-of-core algorithms
- Block-sparse tensor algebra:
  - Irregular data and workload: Load balancing → task-based programming model
- Resolution-adaptive block-sparse tensor algebra:
  - Dynamic irregular data and workload: Load balancing, dynamic data redistribution

Performance Portability Strategy

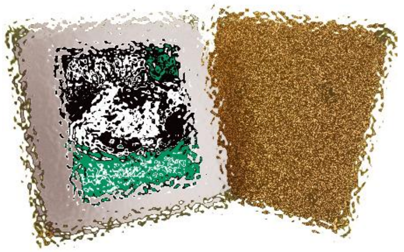
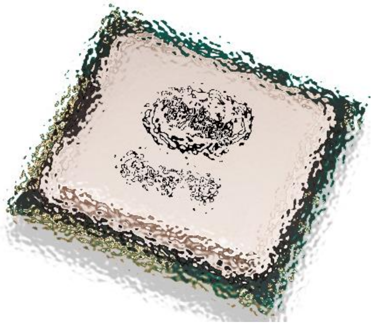
- **Abstract computing system**:
  - Distributed (weakly-coupled) level: The computing system is composed of compute nodes interconnected via network interfaces in accordance with some topology
  - (Semi-)shared (strongly-coupled) level: Each node is composed of multiple compute devices of the same or different kinds, possibly sharing the same (multi-level) memory
- Algorithms are formulated for this abstract computer
- The hardware specificity is masked by driver libraries that provide a device-unified API interface for a set of necessary domain-specific primitives
- New hardware = New driver library, done!

Node-Level Virtualization: Hiding Hardware

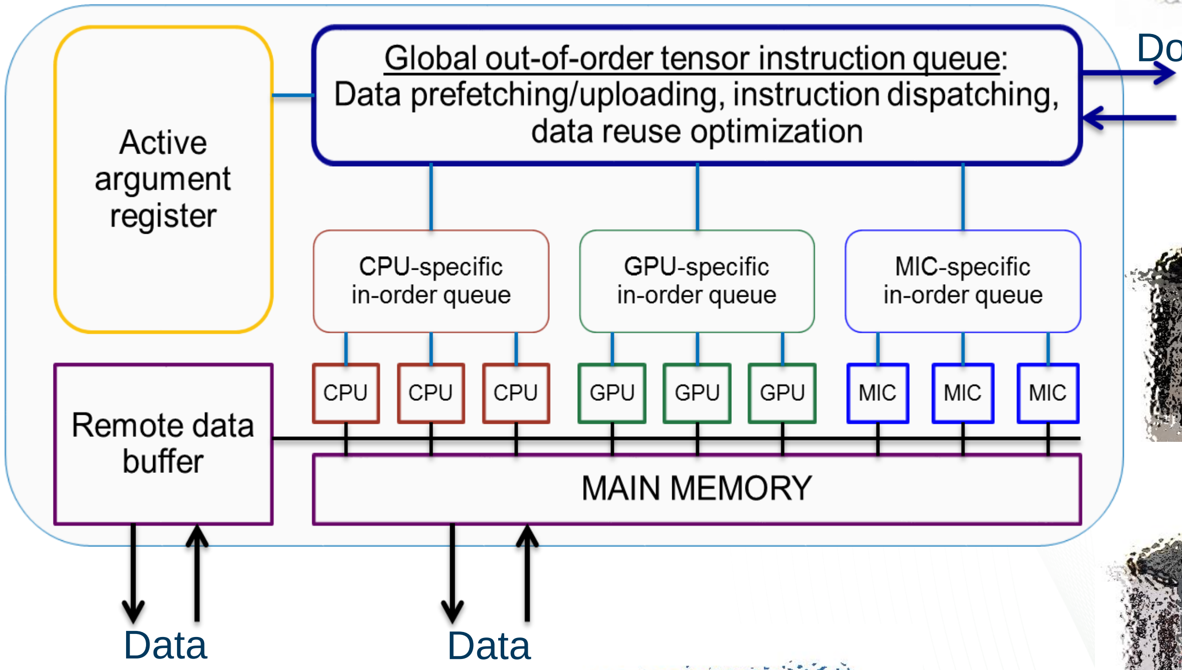
Hardware Processors (CPU, GPU, etc.)

Node-Level Virtualization: Hiding Hardware

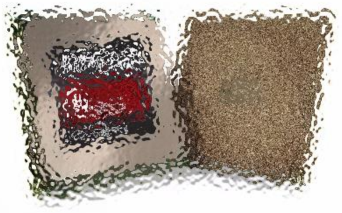
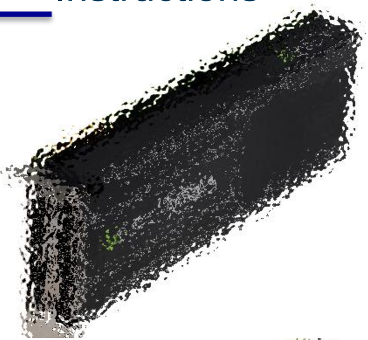
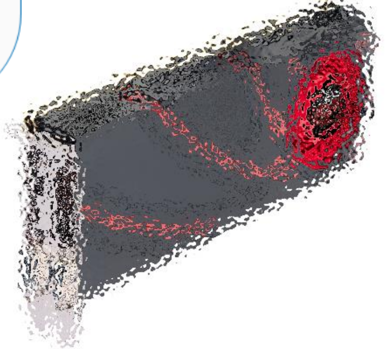
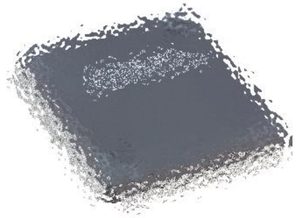
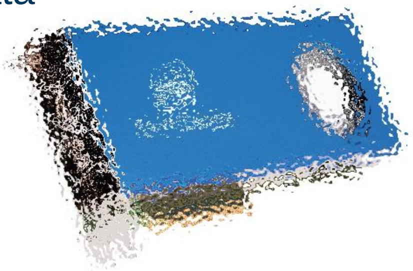
Domain-Specific Virtual Processor

ExaTensor Virtual Processor

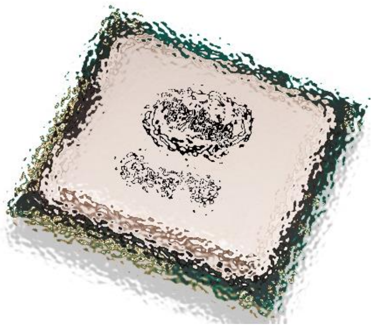


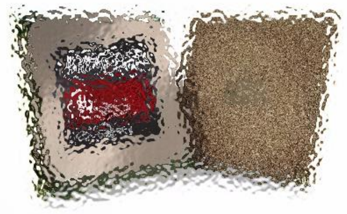
Active argument register
Global out-of-order tensor instruction queue:
Data prefetching/uploading, instruction dispatching, data reuse optimization
Domain-Specific Instructions
CPU-specific in-order queue
GPU-specific in-order queue
MIC-specific in-order queue
CPU
CPU
CPU
GPU
GPU
GPU
MIC
MIC
MIC
Remote data buffer
MAIN MEMORY
Data
Data

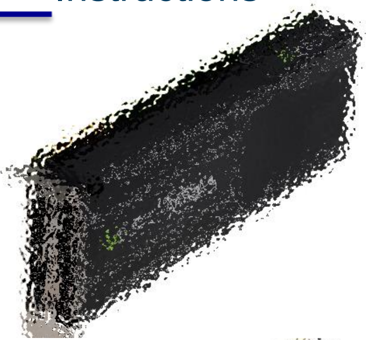






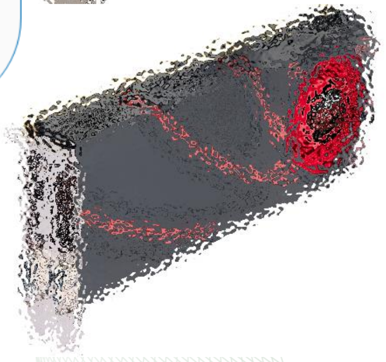
Node-Level Virtualization: Hiding Hardware

Domain-Specific Virtual Processor

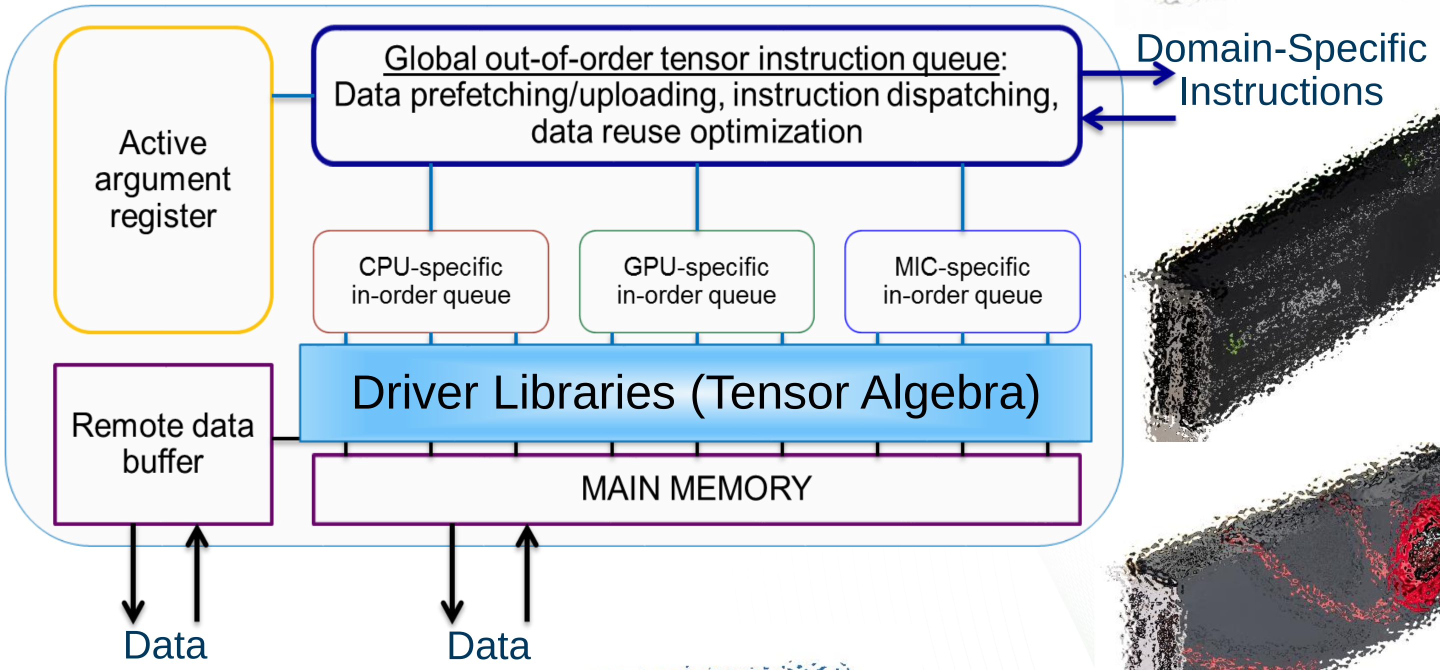


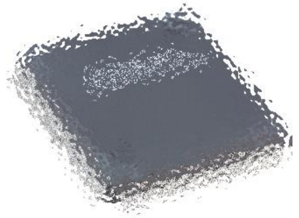




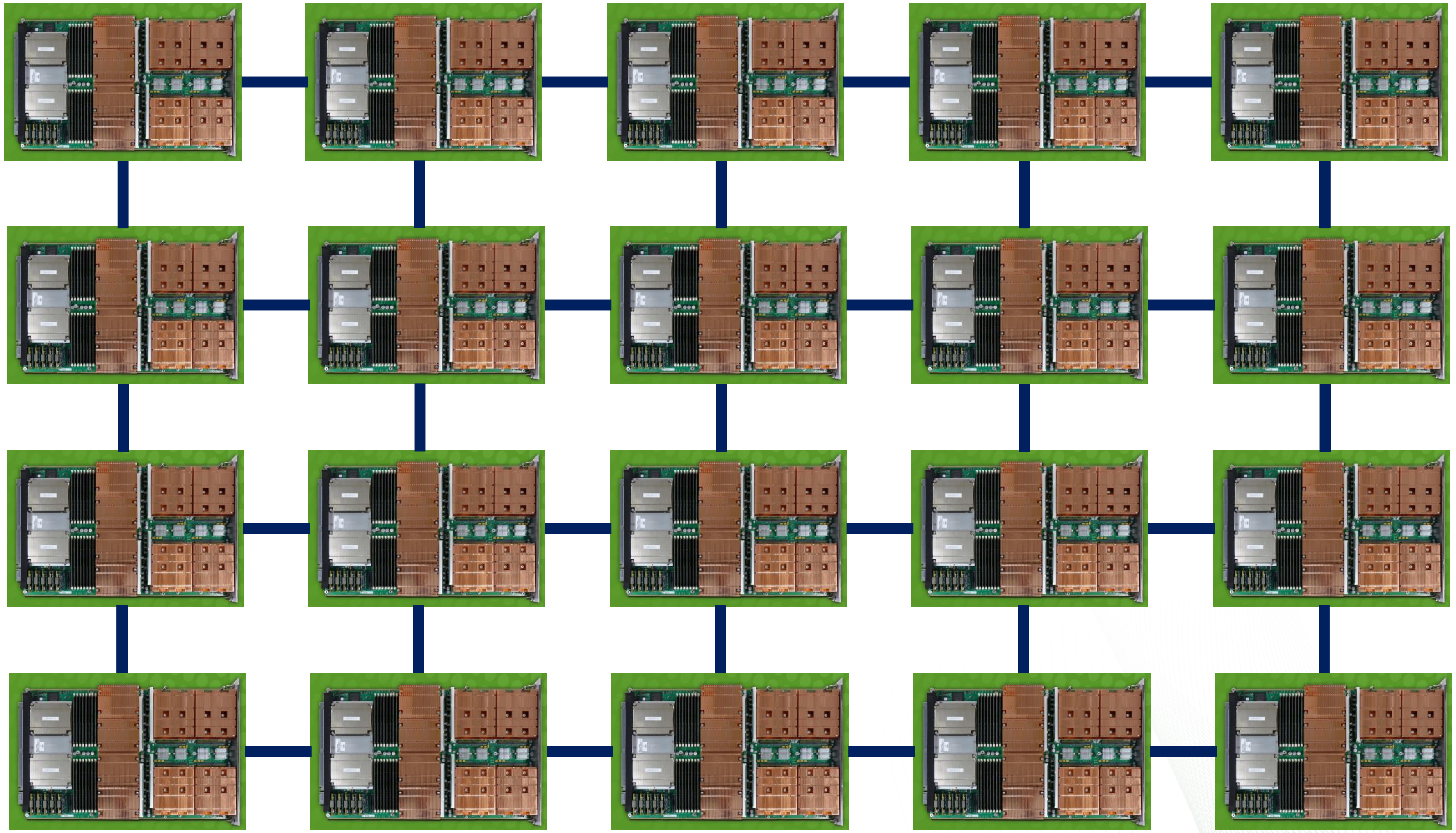


ExaTensor Virtual Processor

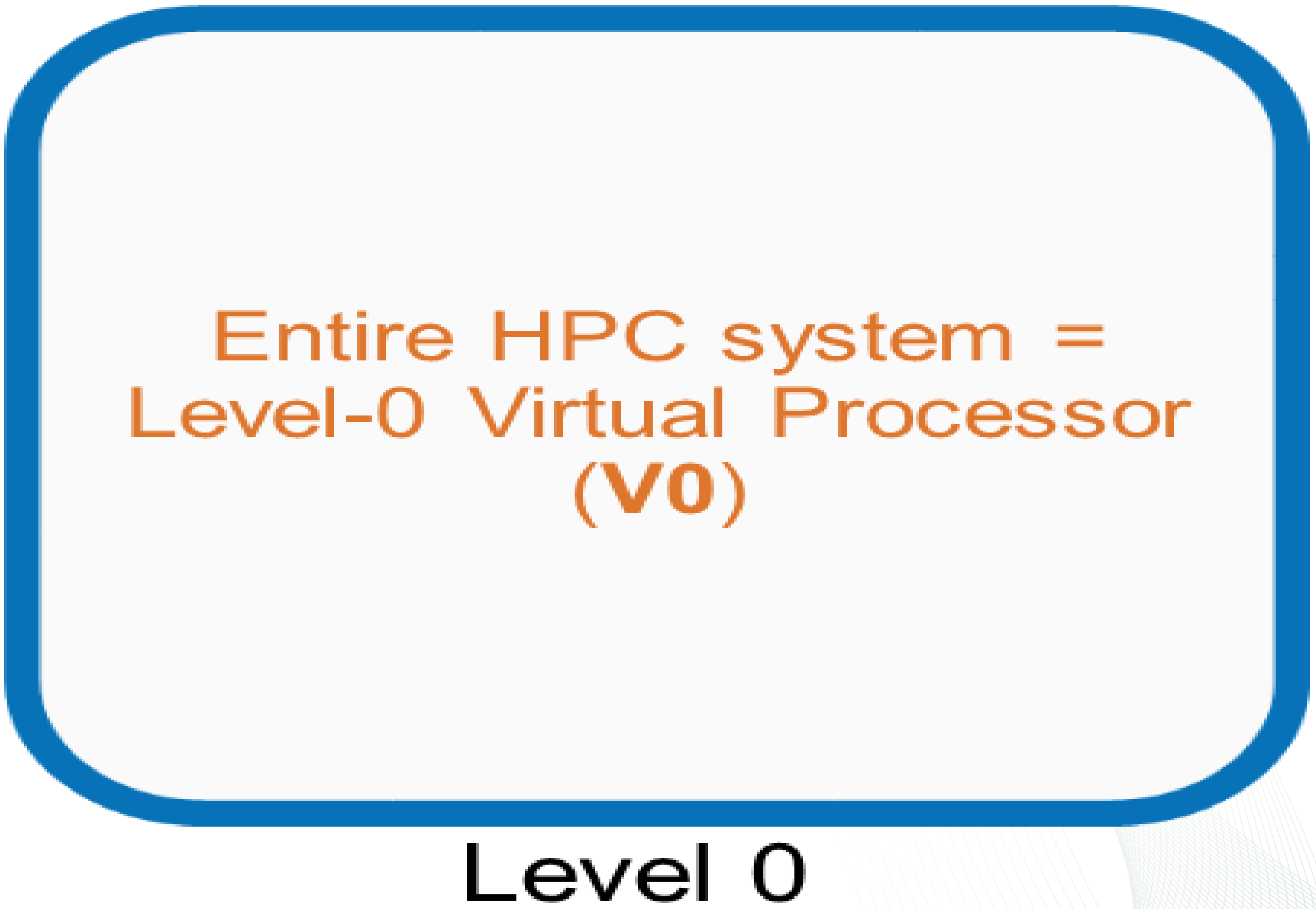

Active argument register
Global out-of-order tensor instruction queue:
Data prefetching/uploading, instruction dispatching, data reuse optimization
Domain-Specific Instructions
CPU-specific in-order queue
GPU-specific in-order queue
MIC-specific in-order queue
Driver Libraries (Tensor Algebra)
Remote data buffer
MAIN MEMORY
Data
Data



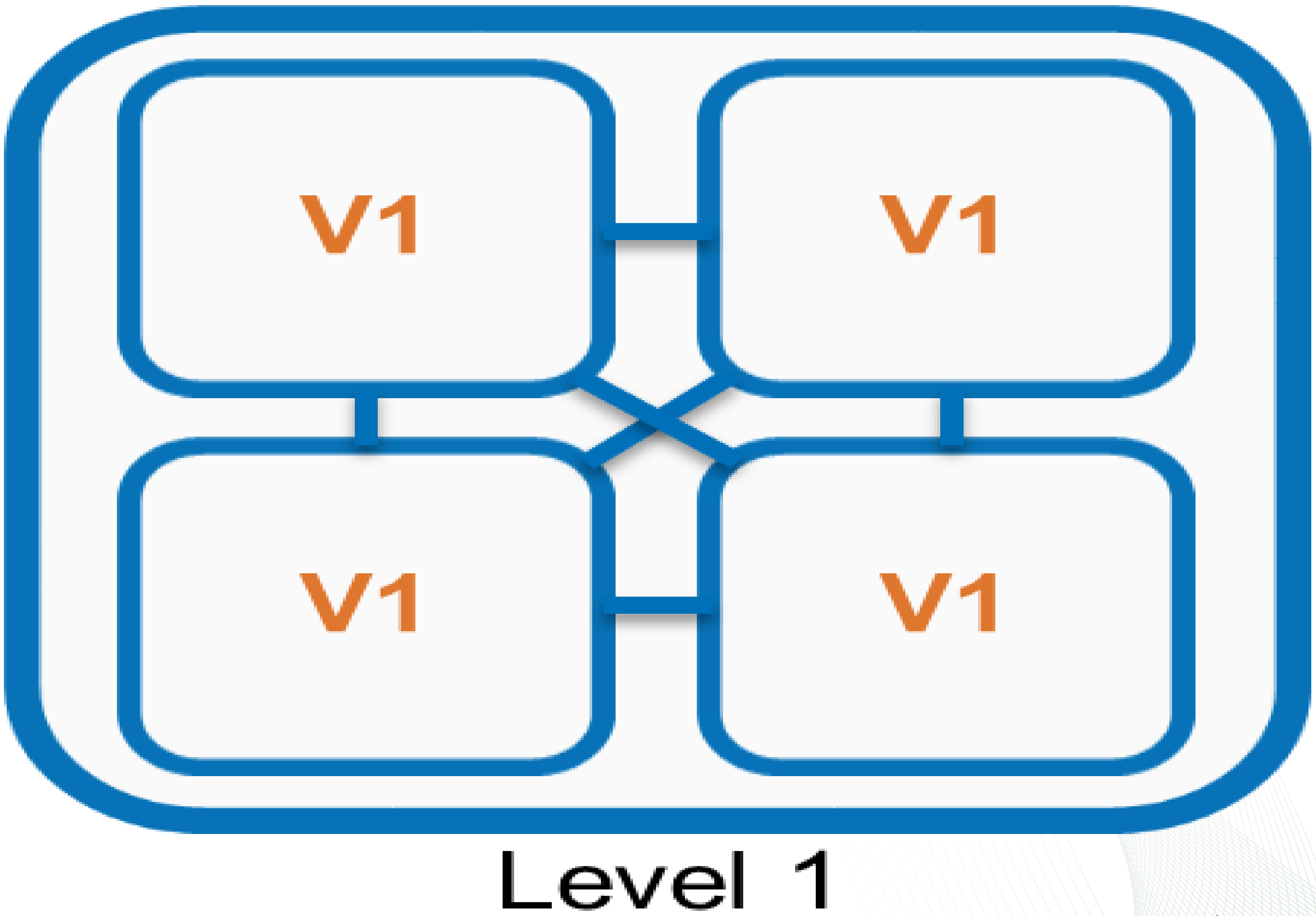
Global Virtualization: Hiding HPC Scale



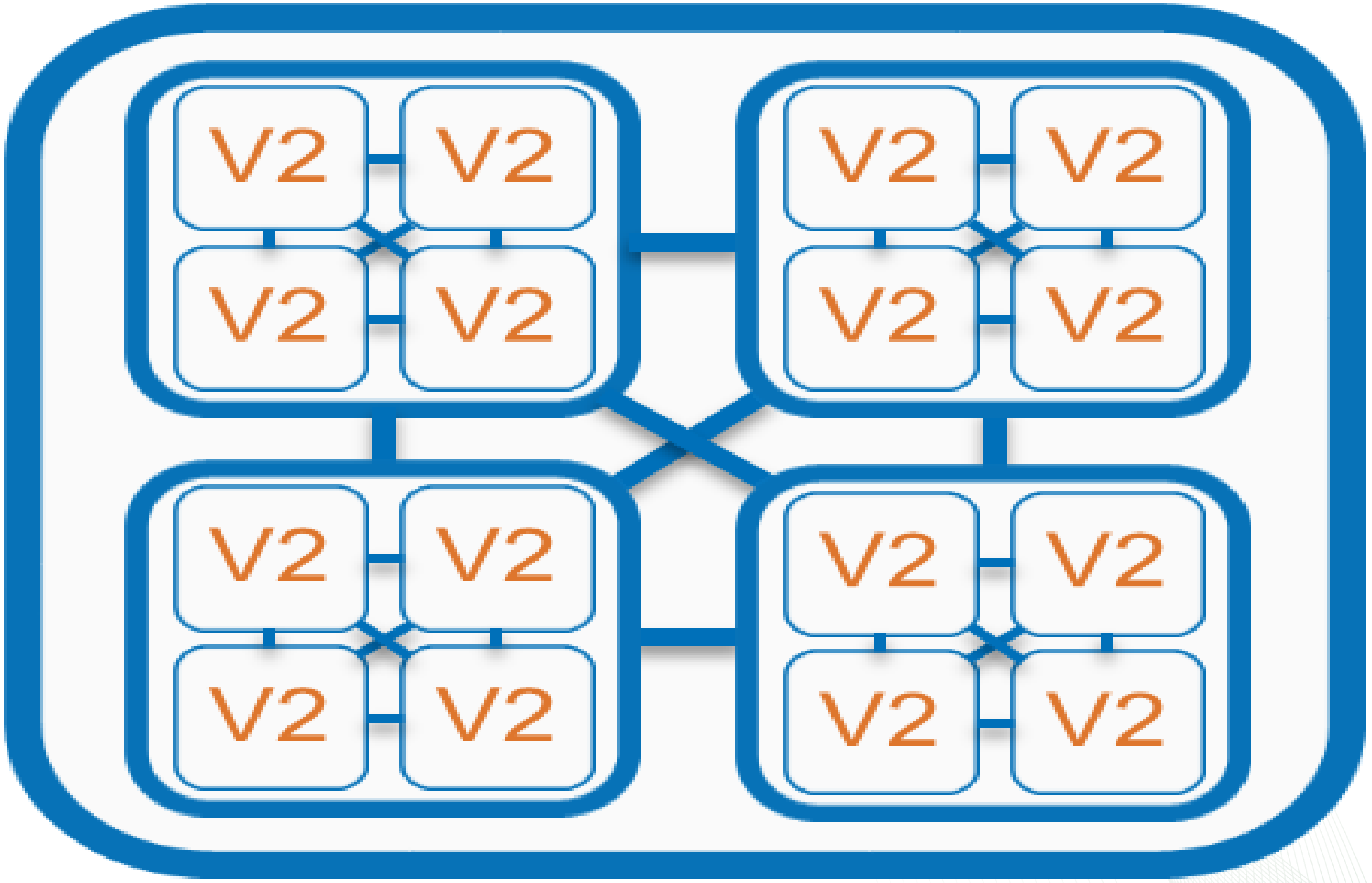
Global Virtualization: Hiding HPC Scale



Global Virtualization: Hiding HPC Scale

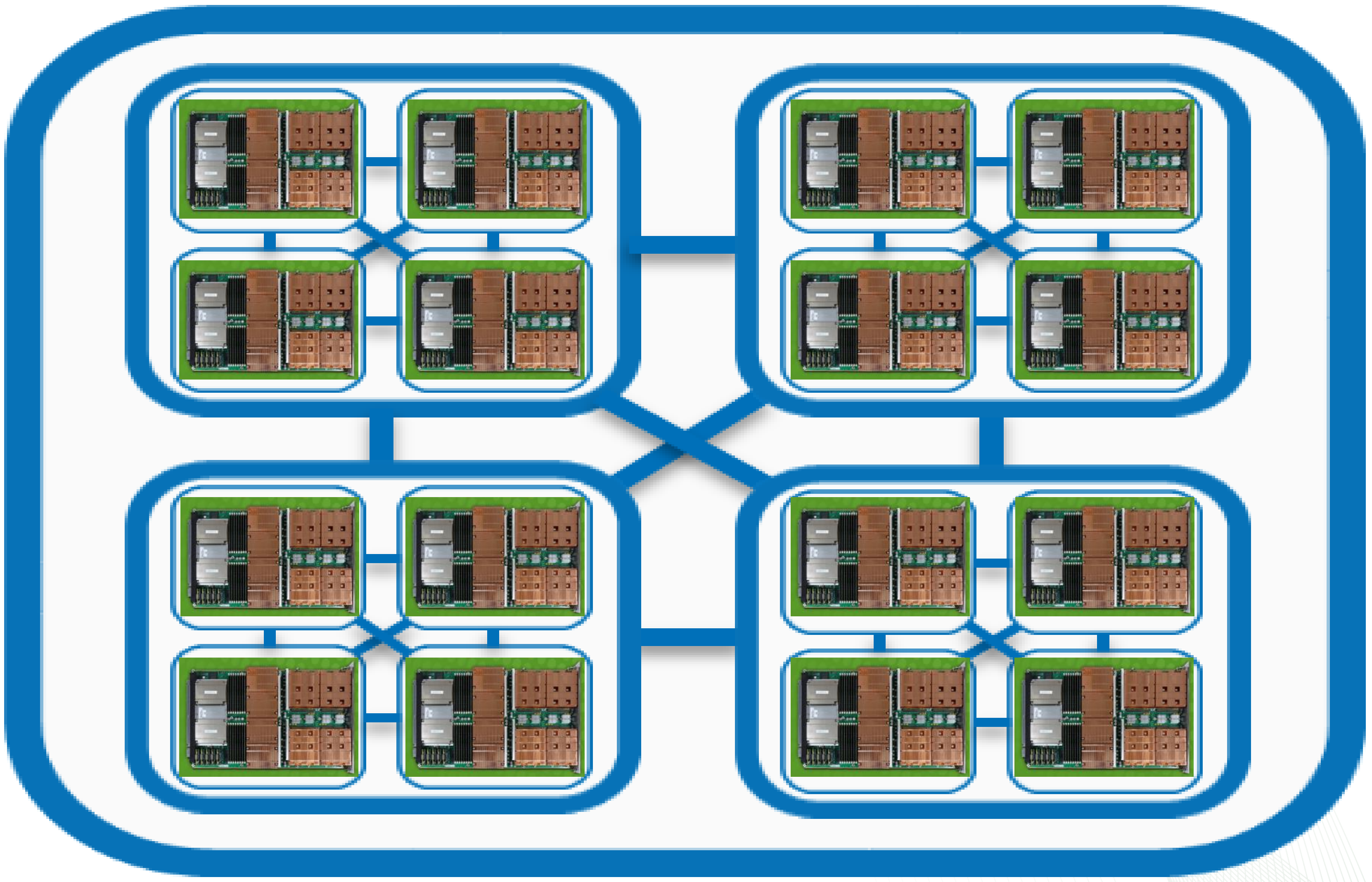


Global Virtualization: Hiding HPC Scale



Level 2

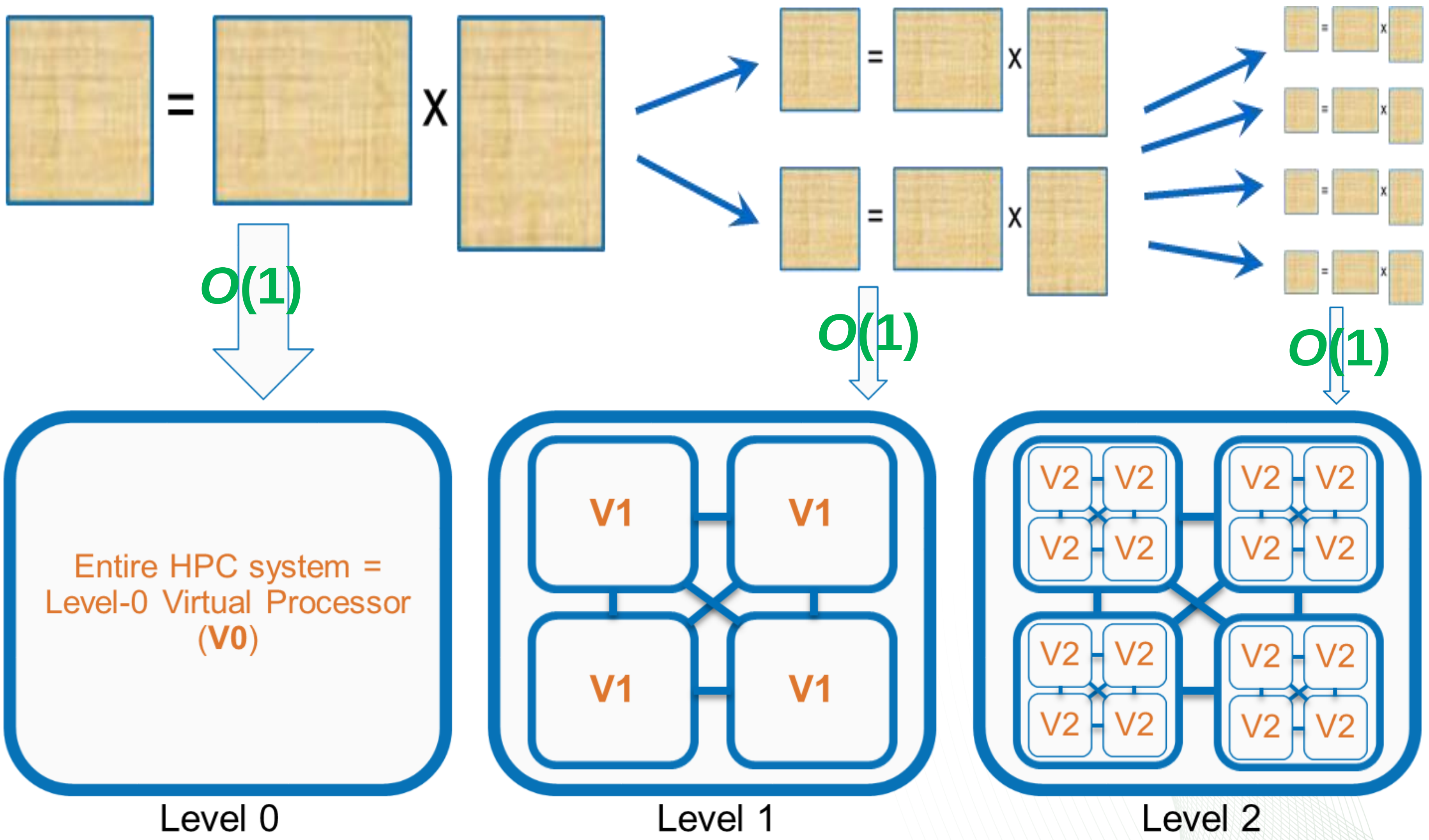
Global Virtualization: Hiding HPC Scale



Level 2

Recursive Dynamic Task Scheduling

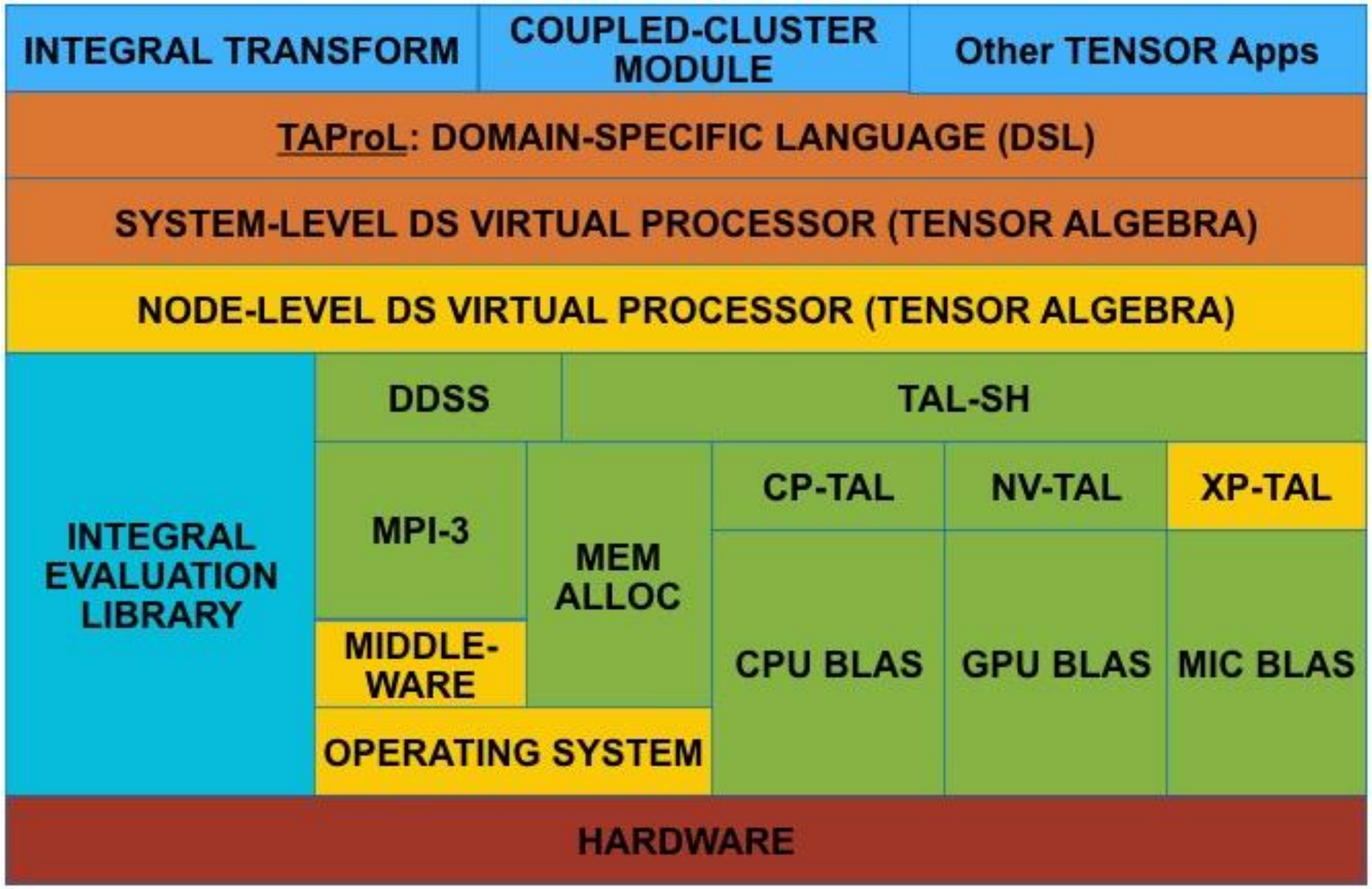
Data storage granularity is decoupled from the task granularity



Oak Ridge National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

ExaTENSOR Software Layers

- Multi-component, library-based hierarchical framework for scalable tensor algebra, portable across different HPC platform: (CPU+ACC, Self-Hosted, …. Future?)


INTEGRAL TRANSFORM
COUPLED-CLUSTER MODULE
Other TENSOR Apps
TAProL: DOMAIN-SPECIFIC LANGUAGE (DSL)
SYSTEM-LEVEL DS VIRTUAL PROCESSOR (TENSOR ALGEBRA)
NODE-LEVEL DS VIRTUAL PROCESSOR (TENSOR ALGEBRA)
INTEGRAL EVALUATION LIBRARY
DDSS
TAL-SH
MPI-3
MEM ALLOC
CP-TAL
NV-TAL
XP-TAL
MIDDLE-WARE
CPU BLAS
GPU BLAS
MIC BLAS
OPERATING SYSTEM
HARDWARE

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

TAL-SH: On-node Tensor Algebra Layer

- Basic tensor algebra operations: Tensor contraction, tensor product, tensor addition, tensor norm, etc.
- Supports multicore CPU and (multiple) NVIDA GPU
- Asynchronous execution on accelerators
- Automatic data transfer management and data transfer pipelining
- User-controlled data presence management
- Tensor contractions:
  - Up to 1.1 TFlop/s double precision on NVIDIA Kepler K20x for arithmetic intensities > 1000
  - Up to 4.1 TFlop/s double precision on NVIDIA Pascal P100 for arithmetic intensities > 1000

(Virtual) Conclusions

- Scalable, performance-portable tensor algebra via hardware and HPC scale abstraction:
  - Hardware agnostic by design;
  - HPC scale "agnostic" by design (still may require communication optimization);
  - Better fault-tolerance (virtual nodes can be turned on/off);
  - A good model for hardware/software co-design;
  - Easy to program, based on a domain-specific higher-level algorithm specification;
  - Better debugging (in terms of higher-level operations);
  - Better profiling (in terms of higher-level operations).
- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract No. DE-AC05-00OR22725.