

# Graph Contractions for Calculating Correlation Functions in Lattice QCD

***Jie Chen, Robert G. Edwards***

Scientific Computing Dept.

Theory Center

Jefferson Lab

Newport News, VA 23606

***Weizhen Mao***

Dept. of Computer Science

William & Mary

Williamsburg, VA 23187



Monday, June 26, 2023

*Jefferson Lab's accelerator site*

# Organization of this Talk

---

- Introduction and Motivation
  - Jefferson Lab
  - Lattice QCD
- Graphs and Graph Contractions
  - Correlation Function and Graphs
  - Graph Classification
  - Graph Contraction
  - Algorithms
- Software Components
  - *Redstar* and *Hadron*
- Performance
- Conclusion

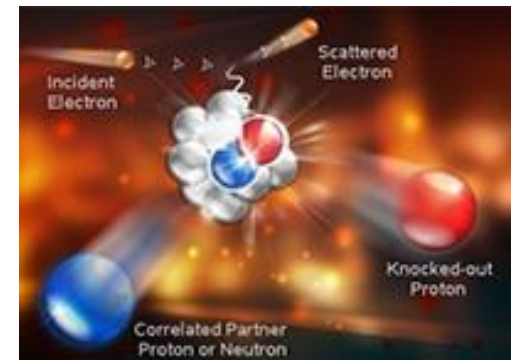
# Jefferson Lab *Exploring the Nature of Matter*

Thomas Jefferson National Accelerator Facility (**Jefferson Lab**) is a U.S. Department of Energy Office of Science national laboratory. Its primary mission is to enable basic research for building a comprehensive understanding of the atom's nucleus by scientists and students worldwide. In addition, the laboratory capitalizes on its unique technologies and expertise to perform advanced computing and applied research with industry and university partners.



*Quarks bound together  
By the strong force*

The protons and neutrons are assembled from quarks and gluons. Jefferson Lab is home to one of the most powerful microscopes in the world for studying these subatomic building blocks: the Continuous Electron Beam Accelerator Facility (**CEBAF**).



# CEBAF AT JEFFERSON LAB

Jefferson Lab's Continuous Electron Beam Accelerator Facility (CEBAF) enables world-class fundamental research of the atom's nucleus. Like a giant microscope, it allows scientists to "see" things a million times smaller than an atom.



## 1 INJECTOR

The injector produces electron beams for experiments.



## 2 LINEAR ACCELERATOR

The straight portions of CEBAF, the linacs, each have 25 sections of accelerator called cryomodules. Electrons travel up to 5.5 passes through the linacs to reach 12 GeV.



## 3 CENTRAL HELIUM LIQUEFIER

The Central Helium Liquefier keeps the accelerator cavities at -456 degrees Fahrenheit.



## 4 RECIRCULATION MAGNETS

Quadrupole and dipole magnets in the tunnel focus and steer the beam as it passes through each arc.



## 5 EXPERIMENTAL HALL A

Hall A is configured with two High Resolution Spectrometers for precise measurements of the inner structure of nuclei. The hall is also used for one-of-a-kind, large-installation experiments.



## 6 EXPERIMENTAL HALL B

The CEBAF Large Acceptance Spectrometer surrounds the target, permitting researchers to measure simultaneously many different reactions over a broad range of angles.



## 7 EXPERIMENTAL HALL C

The Super High Momentum Spectrometer and the High Momentum Spectrometer make precise measurements of the inner structure of protons and nuclei at high beam energy and current.

AT JEFFERSON LAB, NUCLEAR PHYSICISTS STUDY FOUR FUNDAMENTAL AREAS:

- **Quark Confinement** – Addressing one of the great mysteries of modern physics – why quarks only exist together, and never alone.
- **Tests of the Standard Model** – Studying the limits of the Standard Model, a theory that describes the fundamental particles and their interactions.
- **The Physics of Nuclei** – Illuminating the role of quarks in the structure and properties of atomic nuclei, and how these quarks interact with a dense nuclear medium.
- **The Fundamental Structure of Protons and Neutrons** – Mapping in detail the distributions of quarks in space and momentum, culminating in a picture of the internal structures of protons and neutrons.



## 8 EXPERIMENTAL HALL D

Hall D is configured with a superconducting solenoid magnet and associated detector systems that are used to study the strong force that binds quarks together.

# CEBAF AT JEFFERSON LAB

Jefferson Lab's Continuous Electron Beam Accelerator Facility (CEBAF) enables world-class fundamental research of the atom's nucleus. Like a giant microscope, it allows scientists to "see" things a million times smaller than an atom.



AT JEFFERSON LAB, NUCLEAR PHYSICISTS STUDY FOUR FUNDAMENTAL AREAS:

- Quark Confinement – Addressing one of the great mysteries of modern physics – why quarks only exist together, and never alone.
- Tests of the Standard Model – Studying the limits of the Standard Model, a theory that describes the fundamental particles and their interactions.
- The Physics of Nuclei – Illuminating the role of quarks in the structure and properties of atomic nuclei, and how these quarks interact with a dense nuclear medium.
- The Fundamental Structure of Protons and Neutrons – Mapping in detail the distributions of quarks in space and momentum, culminating in a picture of the internal structures of protons and neutrons.



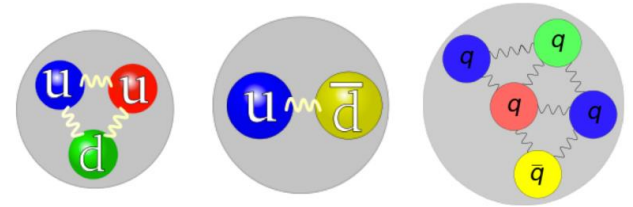
## AT JEFFERSON LAB, NUCLEAR PHYSICISTS STUDY FOUR FUNDAMENTAL AREAS:

- Quark Confinement – Addressing one of the great mysteries of modern physics – why quarks only exist together, and never alone.
- Tests of the Standard Model – Studying the limits of the Standard Model, a theory that describes the fundamental particles and their interactions.
- The Physics of Nuclei – Illuminating the role of quarks in the structure and properties of atomic nuclei, and how these quarks interact with a dense nuclear medium.
- The Fundamental Structure of Protons and Neutrons – Mapping in detail the distributions of quarks in space and momentum, culminating in a picture of the internal structures of protons and neutrons.

# Quantum Chromodynamics (QCD)

- Why study QCD?

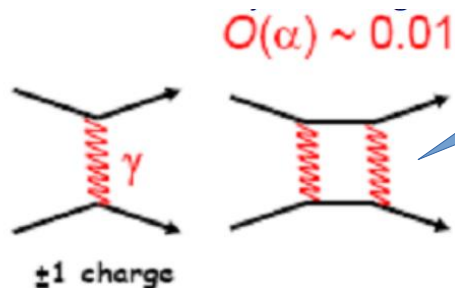
- Standard Model: Strong Interaction
- Make composite “stuff”: Baryons, Mesons ...
  - [Spectroscopy, Energy, Structure]
- To interpret experimental results.



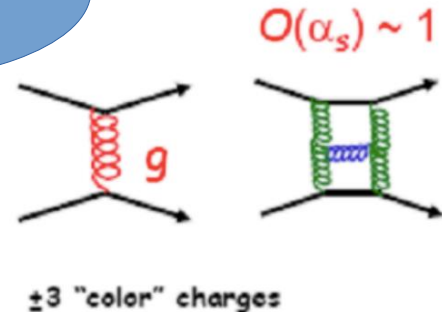
- QCD

- Gluons: force carrier, 100 times stronger than electromagnetism
  - Strong coupling constant:  $\alpha_s \sim 1$
- Quarks:  $u, d, s, c, b, t$  (red, green, blue)

- QCD versus QED



Perturbation Method



Non-perturbation Method

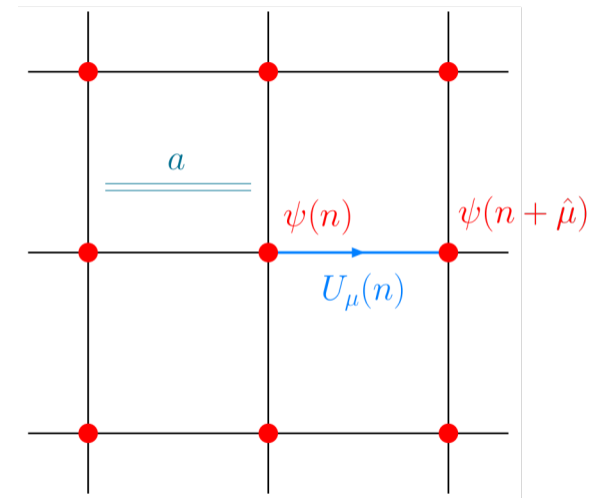
# Lattice Quantum Chromodynamics (LQCD)

- Ken Wilson developed lattice QCD to go beyond perturbation theory.

- continuum is replaced by a 4-dim grid of space-time points
- quarks described by complex fields with 3 or 3X4 components
- gluons 3X3 complex unitary matrices

- Three Stages of Calculations.

- Generation of ensembles of gauge fields.
  - Supercomputers
- Creation of quark propagators.
  - Solutions of Dirac operator (matrix inversions)
- Calculating correlation functions.
  - **Wick contractions** on many gauge configurations (statistical accuracy).
    - Expensive in computing resource ~ first step.



# Correlation Functions in LQCD

- To deduce the quantities of interest known as physics observables such as the energy spectrum of hadrons.

- 2-pt correlation function.  $C(t', t) = \langle \chi(t') \chi^\dagger(t) \rangle$

$$\chi = \sum W_h^{a_1, \dots, a_{n_q}} q(a_1) q(a_2) \dots q(a_{n_q}) \quad \longrightarrow \quad \frac{N!}{n_q!(N - n_q)!}$$

$$\begin{aligned} C(t', t) &= \langle \chi(t') \chi^\dagger(t) \rangle \\ &= \sum_{(\vec{a}, \vec{a}')} W_{a'_1, \dots, a'_{n_q}}^{a_1, \dots, a_{n_q}} \langle q(a_1) \dots q(a_{n_q}) \bar{q}(a'_1) \dots \bar{q}(a'_{n_q}) \rangle \end{aligned}$$

$$C(t', t) = \left( \prod_{l=1}^{n_q-1} (-1)^l \right) \sum_{\vec{k} \in S_{n_q}} \epsilon^{k_1, \dots, k_{n_q}} S_{1, k_1} \dots S_{n_q, k_{n_q}} \quad \longleftarrow \quad \text{Wick Contractions}$$

$$S_{i,j} := S(a_i; a'_j) = q_i \bar{q}_j \quad \longleftarrow \quad \text{Quark Propagator}$$

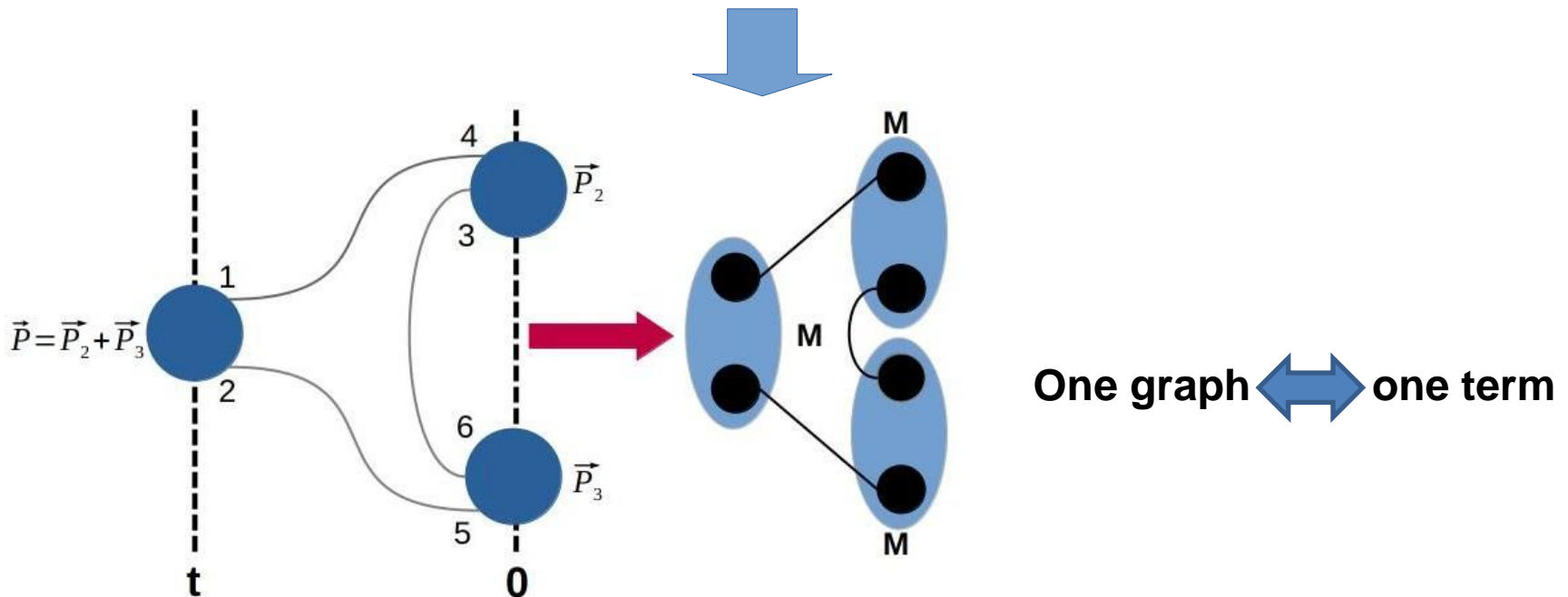


# Redstar: Correlation Functions and Graphs

$$C(t', t) = \langle \chi(t') \chi^\dagger(t) \rangle \quad \hat{H}|k\rangle = E_k|k\rangle \quad C(t, t') = \sum_k |\langle \chi|k\rangle|^2 e^{-E_k(t'-t)}$$

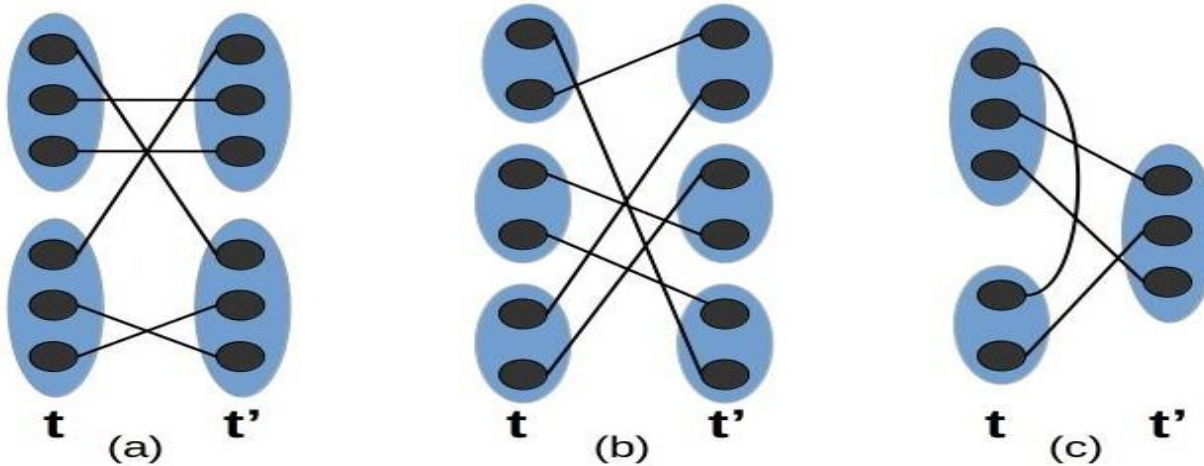
Smearing and distillation

$$C(t, 0) = \sum_{\vec{p}_2, \vec{p}_3 | \vec{p}} c_{\vec{p}_2, \vec{p}_3} M^{12}(\vec{p}, t) P^{25}(t, 0) M^{56}(\vec{p}_2, 0) P^{63}(0, 0) M^{34}(\vec{p}_3, 0) P^{41}(0, t)$$

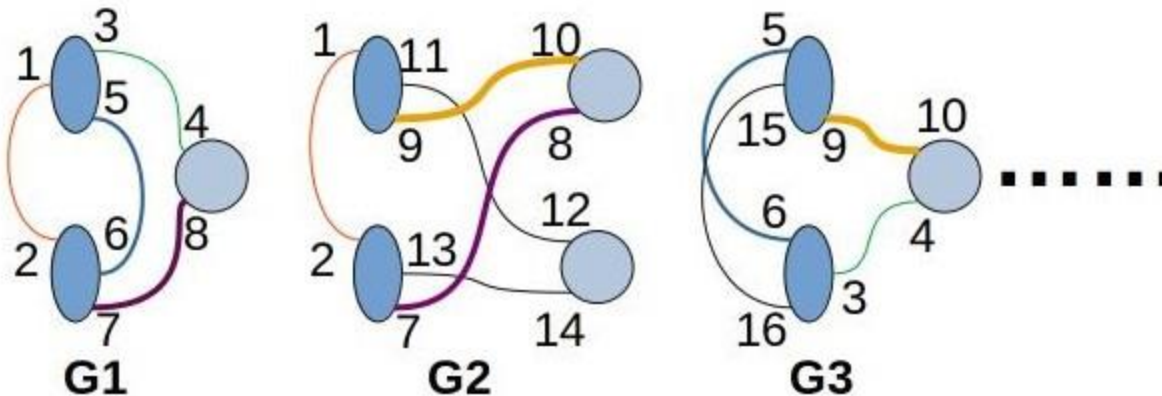


# Graphs, Graphs, Graphs ...


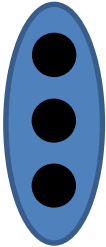
- Different types of graphs



- Different graphs can share the same edges



# Graph (*Hadron*) Nodes (V)

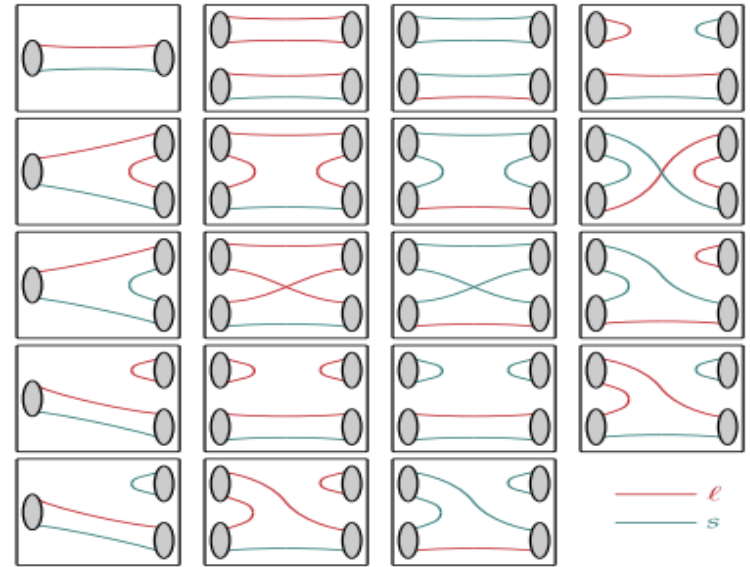
- Meson (2 quarks),  Baryon (3 quarks) 
- Meson
  - Each vertex  $M_{\alpha\beta}^{ij}$  where  $\alpha, \beta = 0, 1, 2, 3$  (spins) and  $i, j \sim$  hundreds (distillation space).
  - e.g. each vertex occupies  $384 * 384 * 16 * 16 = 37$  MB
- Baryon
  - Each vertex  $B_{\alpha\beta\gamma}^{ijk}$  where  $\alpha, \beta, \gamma = 0, 1, 2, 3$  (spins) and  $i, j, k \sim$  100s (distillation space)
  - e.g. each vertex occupies  $128 * 128 * 128 * 16 * 64 = 2$  GB, each vertex occupies about  $128 * 128 * 128 * 16 * 8 = 268$  MB if only two upper spin indices are used ( $\alpha, \beta, \gamma = 0, 1$ ).

# Graph Edges (E)

- Meson-Meson contraction  $A_{\alpha\beta}^{ij} B_{\beta\gamma}^{jk}$ 
  - Batched matrix multiplication with batch size  $\leq 64$
  - $O(N^3)$  for calculations,  $O(N^2)$  for memory
- Meson-Bryon contraction  $A_{\alpha\beta}^{ij} B_{\beta\gamma\delta}^{jkl}$ 
  - Batched tensor contraction with batch size  $\leq 256$
  - $O(N^4)$  calculation complexity,  $O(N^3)$  memory
- Baryon-Baryon Contraction
  - One index contraction  $A_{\alpha\beta\gamma}^{ijk} B_{\gamma\delta\epsilon}^{klm}$ 
    - $O(N^5)$  calculation complexity,  $O(N^4)$  memory
    - Huge memory (e.g  $128^4 * 16 * 16 = 68\text{GB}$  for two upper spins)
    - Batched tensor contraction with batch size up to 1024
  - Two-index contraction  $A_{\alpha\beta\gamma}^{ijk} B_{\gamma\beta\delta}^{jkl}$ 
    - $O(N^4)$  calculation complexity,  $O(N^2)$  memory.

# How Many Graphs?

- *Number of Graphs*  $\propto N!$
- $N$  is the number of freedom of degrees.
- Recently approaching 1M graphs.
- Calculations across many time slices and configurations.

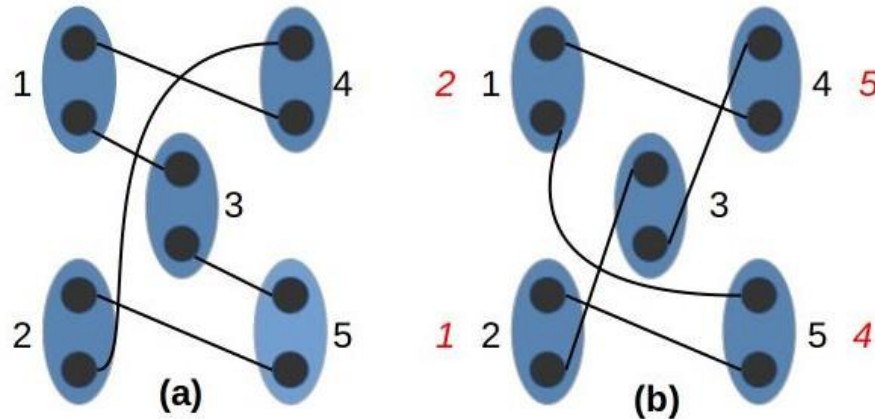


$I=1/2$   $K^*\pi$  arXiv:1406.4158

Name	$a_0$	$f_0$	$a_1$	roper	Deuterium	Tritium
# Graphs	19,041	27,999	385,512	84,894	119,191	6,208
Type	MM	MM	MM	MB	BB	BB
Vector Size	256	256	128	64	64	32

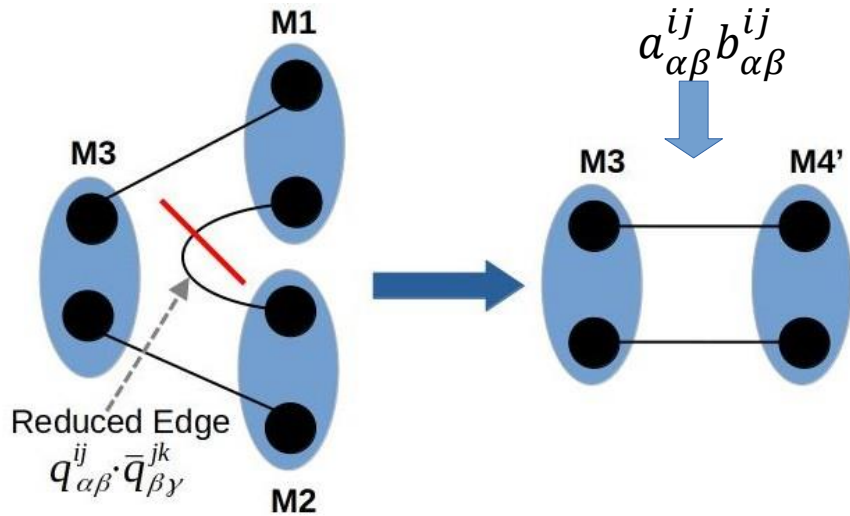
# Graph Classification

- Identify unique graphs among many graphs.

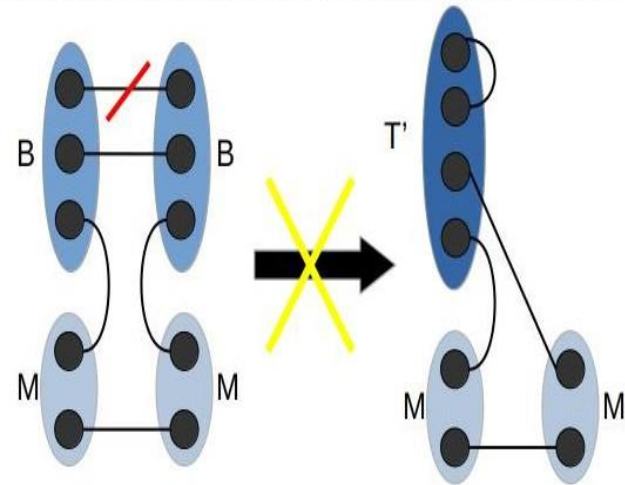
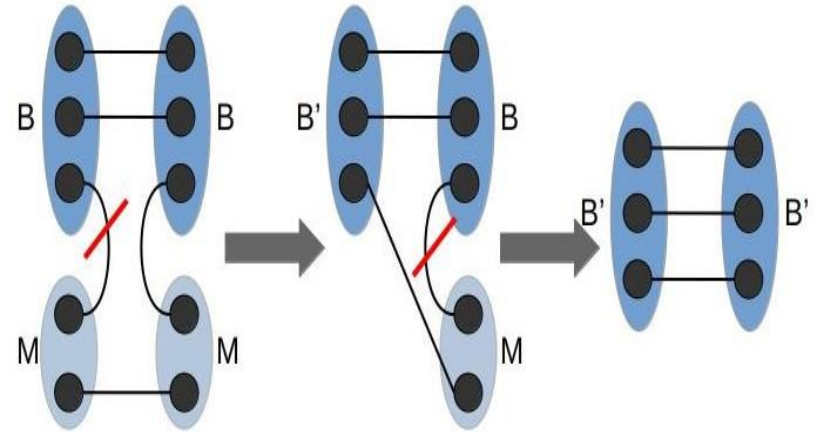
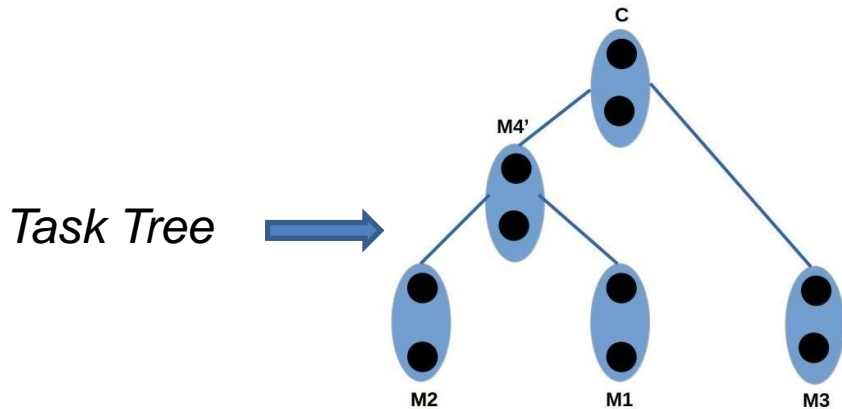


- Naive way takes  $O(n^2)$  steps.
  - For each unclassified graph, generate many graphs through eligible permutations of the nodes. Compare the generated graphs one-by-one with the classified graphs.
- A better way through *Canonical Labeling* using **Nauty**.
  - *Redstar* utilizes canonical labeling for each graph in parallel.
    - Generates a hash value for each labeled graph.
  - $O(n)$  steps to compare all the hash values.

# Graph Contractions



Removal of one edge after another until two hadron nodes are left.



**The sequence of the edge removals is important**

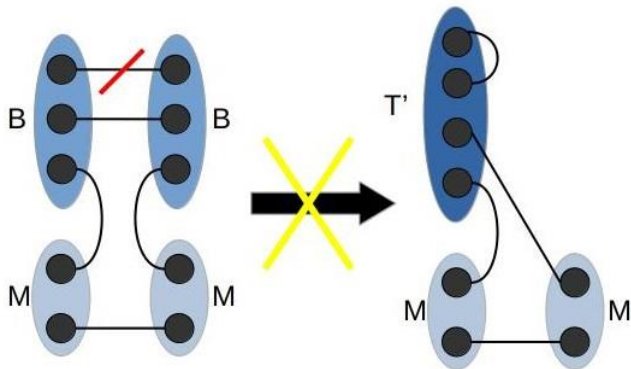
# Graph Contractions

- Many ways to contract a graph

- Which edge to reduce ?
  - A single edge can appear in many graphs:  $E_N$  (# of appearance)
    - Re-usability
  - A single edge removal represents a batched tensor contraction.
    - Computing and memory complexity
- Removal of multiple edges at a time?
  - Single edge removal represents a one-index contraction. Double edge removal stands for two-index contraction.

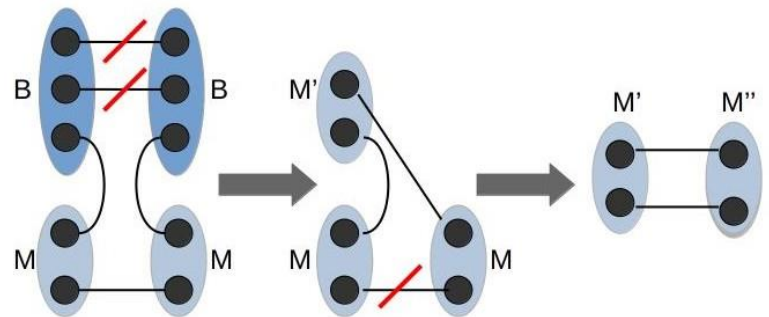
- $O(N^5)$  computing

- $O(N^4)$  memory



- $O(N^4)$  computing

- $O(N^2)$  memory





# Graph Contraction Algorithms

- Naive algorithm

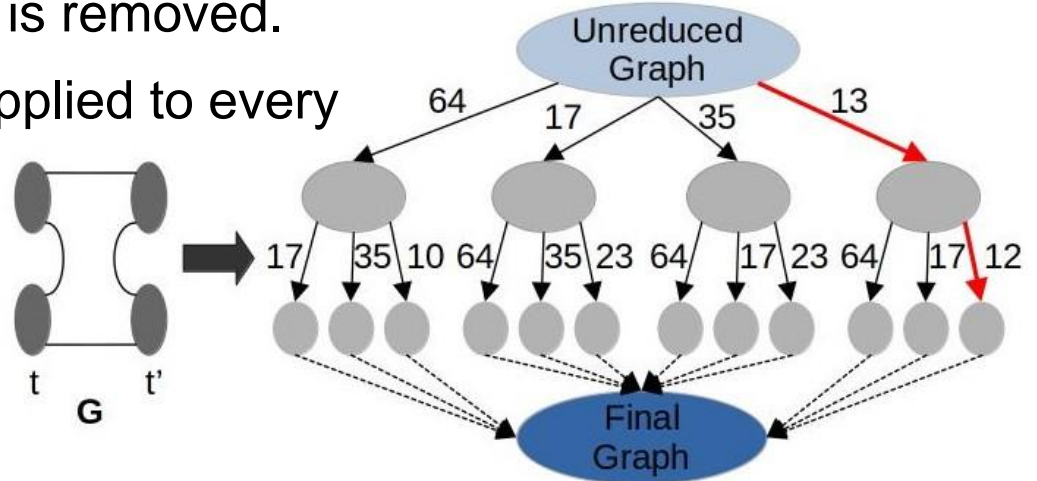
- Eliminate the edges sequentially according to the order in the adjacency matrix representing a graph.

- *Redstar*

- Assign a weight for every edge.  $\frac{C}{E_N}$ 
  - $C$  is the computing complexity of the removal of the edge.  $E_N$  is the number of appearances of this edge in all the graphs.
- A contraction tree is constructed. Each node represents the state of the graph after an edge is removed.

- Dijkstra's algorithm is applied to every tree graph to find the shortest path.

- Minimum spanning tree algorithm is utilized for complicated graphs.

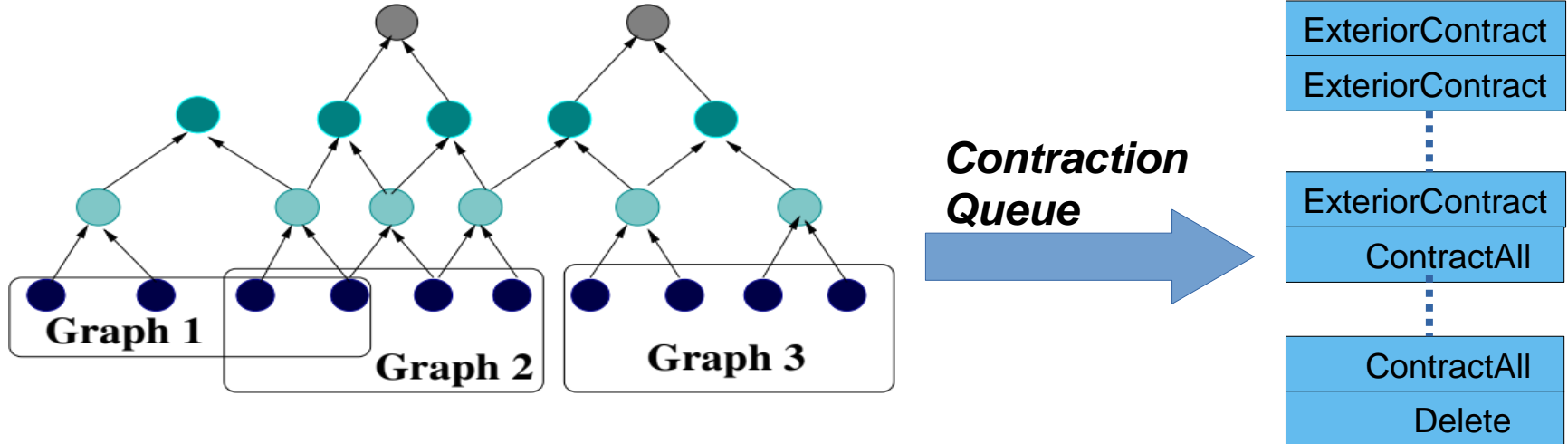


# Optimal Graph Contraction Path

- *Redstar* contraction algorithm reduces the total computing cost.
  - For a multi-meson system, it has the largest total  $E_N$  which yields the highest total computing re-usability and leads to the least total number of contractions for all the graphs.
  - For a baryon system, it eliminates some of the most expensive calculations in addition to promoting re-usability of the edges with high number of appearances in all the graphs.
- The optimal contraction path records the order of the tensor contraction calculations of a graph (*Wick contractions* term).
  - This order is converted into a task tree.

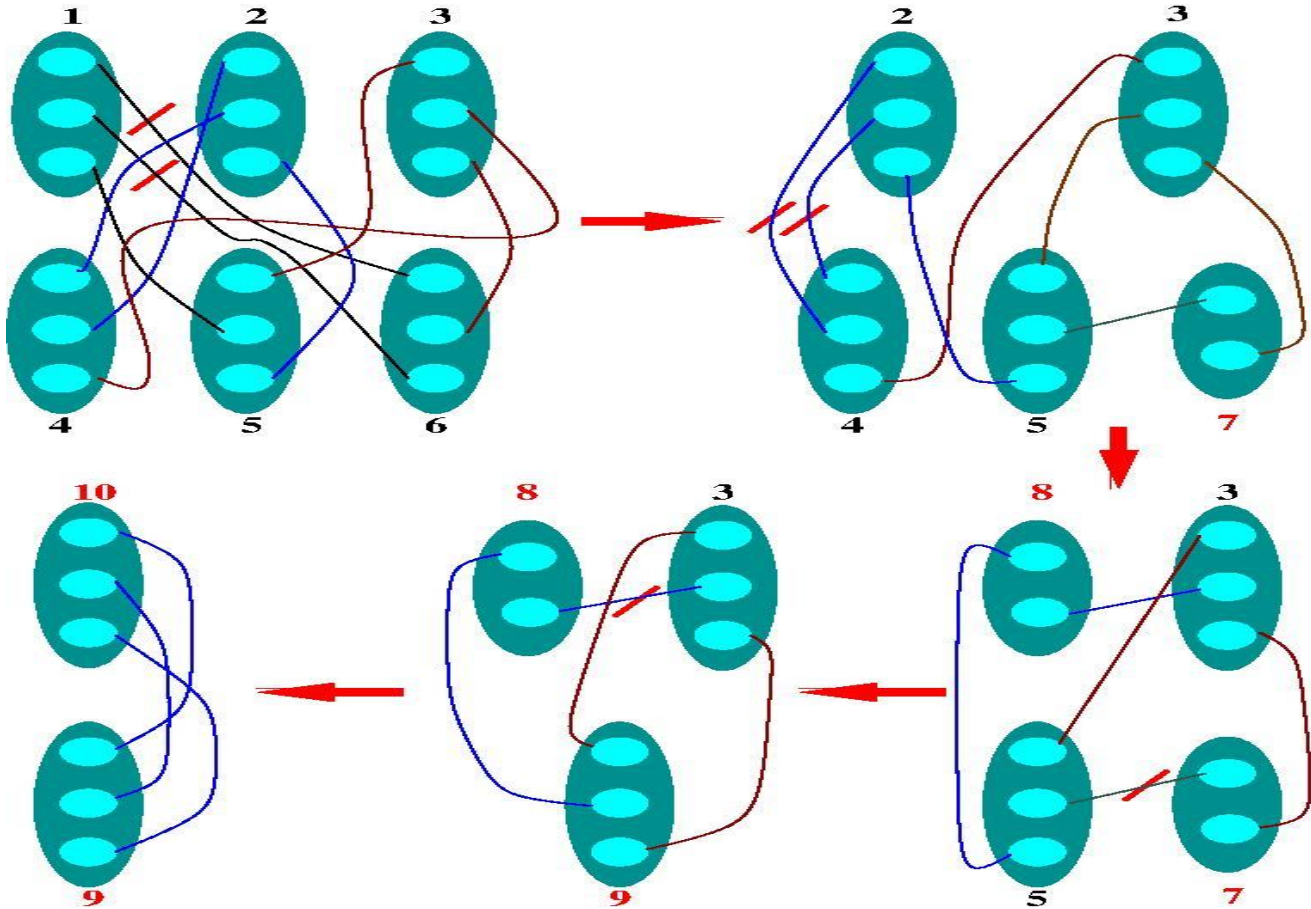
# Reduction in Memory Footprint

- The task trees of all graphs are combined into a large contraction ***evaluation DAG*** for all the graphs.

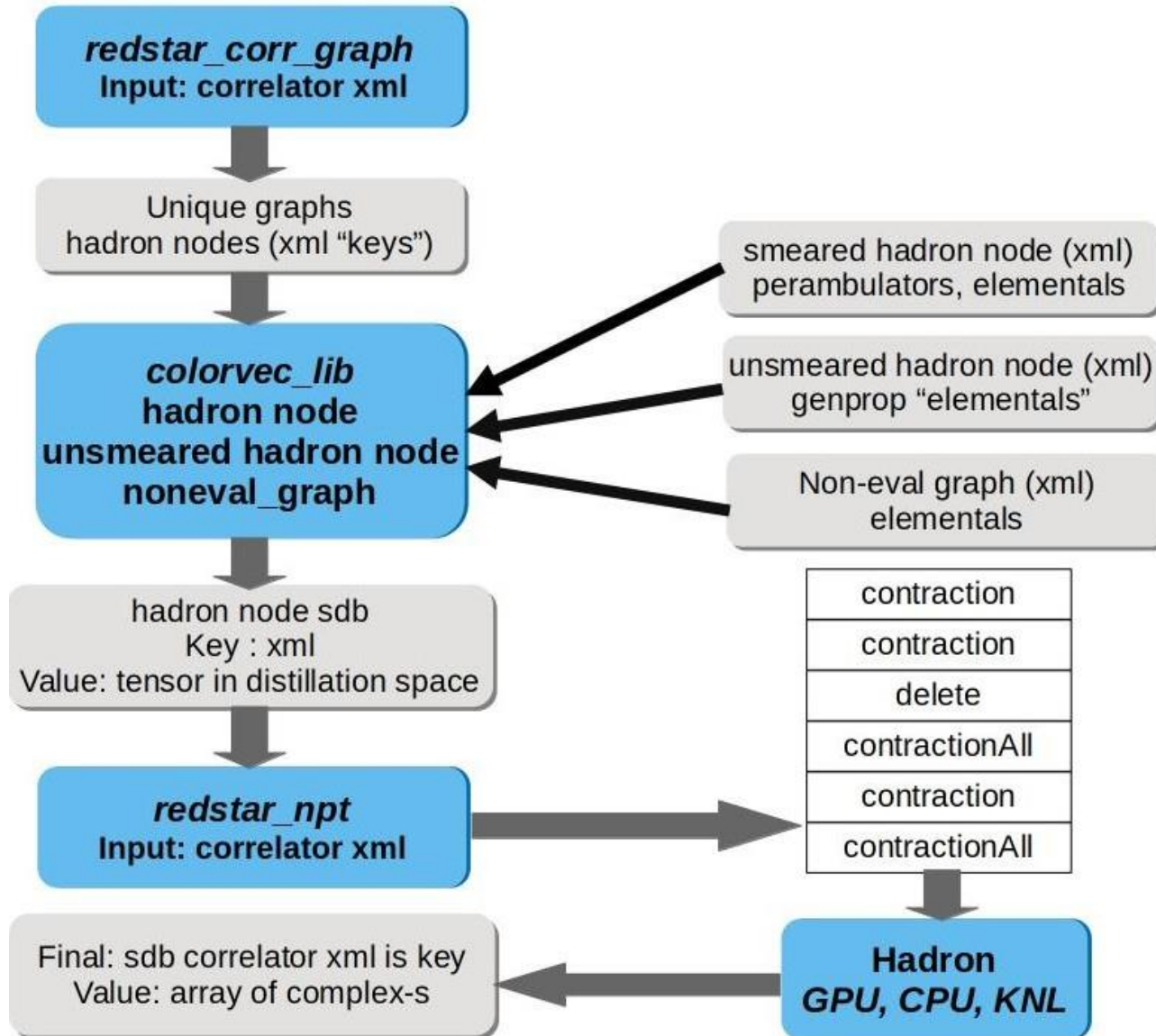


- Redstar* groups similar graphs together when it builds the large DAG to reduce the memory footprint.
  - The result of an edge reduction (temporary hadron object) is cached if it will be used in other contractions from other graphs.
  - The maximum number of cached objects is called High-Water-Mark (***HWM***).
  - Cached objects can be deleted sooner if similar graphs are clustered together. ➡ Smaller ***HWM***.

# Graph Contraction Algorithms



# Redstar Software Suite

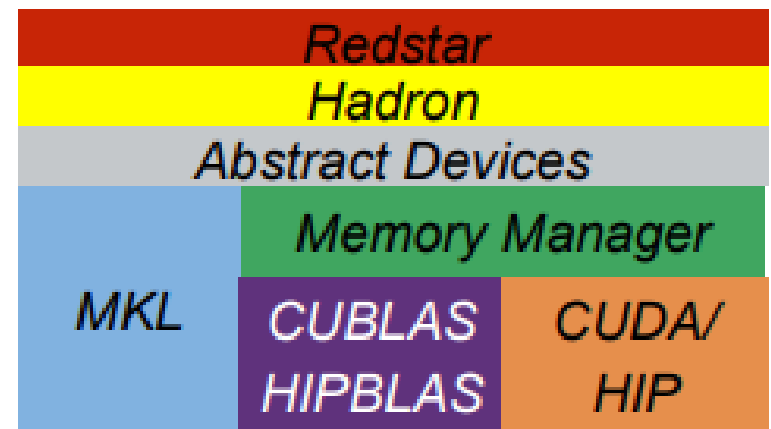


# Hadron Library

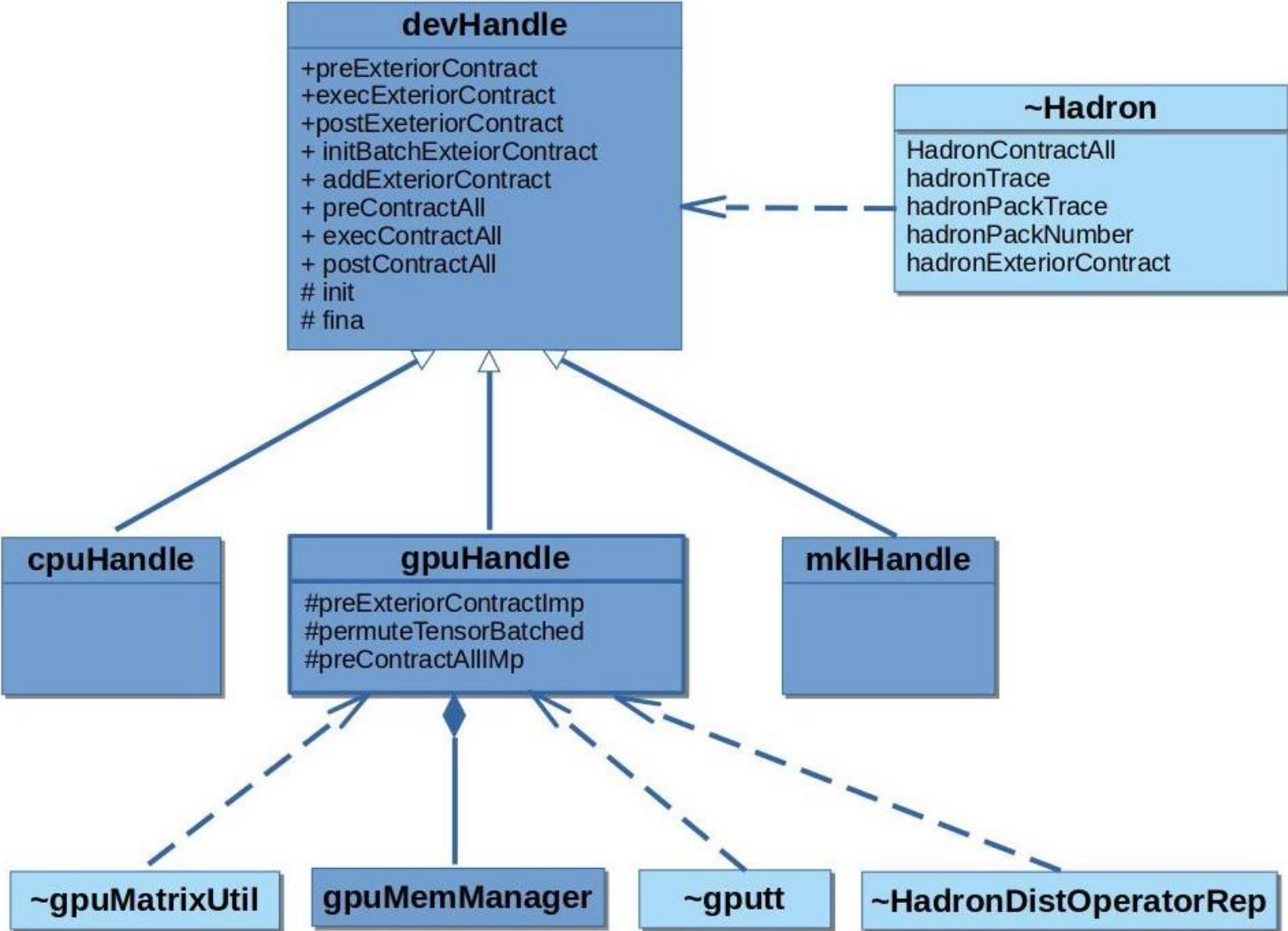
- Handles real batch tensor contractions from the contraction queue generated by *Redstar*.
  - $h_{ijk}^{\alpha\beta\gamma}$  Represents a hadron object.  $\alpha, \beta, \gamma$  are spin [0,3], and  $i, j, k$  are distillation space index (N ~ 100s for baryons,  $100 < N < 1000$  for mesons)
  - Two-index contraction:  $a_{ij}^{\alpha\beta} = b_{ikl}^{\alpha\delta\lambda} \bullet c_{kli}^{\delta\lambda\beta}$
  - One-index contraction:  $a_{ijkm}^{\alpha\beta\gamma\delta} = b_{ijl}^{\alpha\beta\lambda} \bullet c_{lkm}^{\lambda\gamma\delta}$

- Same APIs across different devices.

- C++ device handle base class.
- MKL/Openblas cpu device.
- GPU device.
- *Intel oneAPI (coming soon)*



# Hadron Library C++ Interface



# Hadron Contraction APIs

HadronDistOperatorRep

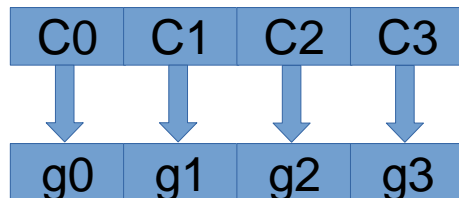
```
hadronExteriorContract(const HadronDistOperatorRep& src1_rep,  
                      int ind_j,  
                      const HadronDistOperatorRep& src2_rep,  
                      int ind_k);
```

HadronDistOperatorRep

```
hadronExteriorContract(const HadronDistOperatorRep& src1_rep,  
                      const std::vector<int>& ind_j,  
                      const HadronDistOperatorRep& src2_rep,  
                      const std::vector<int>& ind_k);
```

std::vector<HadronDistOperatorRep>

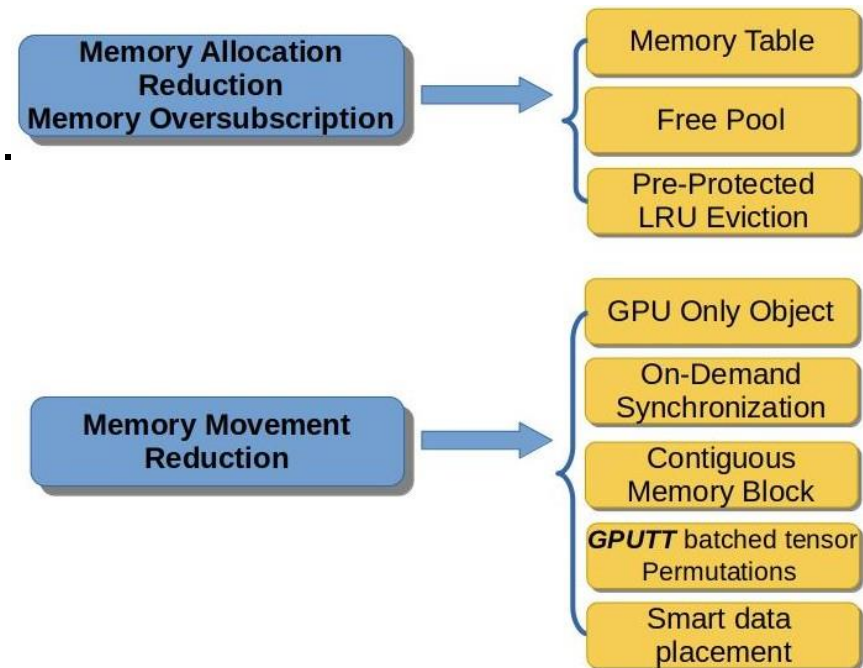
```
hadronExteriorContract(const std::vector<HadronDistOperatorRep>& src1,  
                      const std::vector<std::vector<int>>& ind_j,  
                      const std::vector<HadronDistOperatorRep>& src2,  
                      const std::vector<std::vector<int>>& ind_k);
```





# Hadron Contractions on CPUs and GPUs

- Many-core CPUs:
  - Execute "a group of batched tensor contractions".
  - Fast batched tensor permutation (*mctt*).
- GPUs:
  - Large memory of a hadron object.
  - Data movements between GPUs and CPUs and among GPUs.
  - GPU memory size limitation.
  - Batched tensor permutation on GPUs.



# Performance Evaluation Test Settings

- Correlation functions

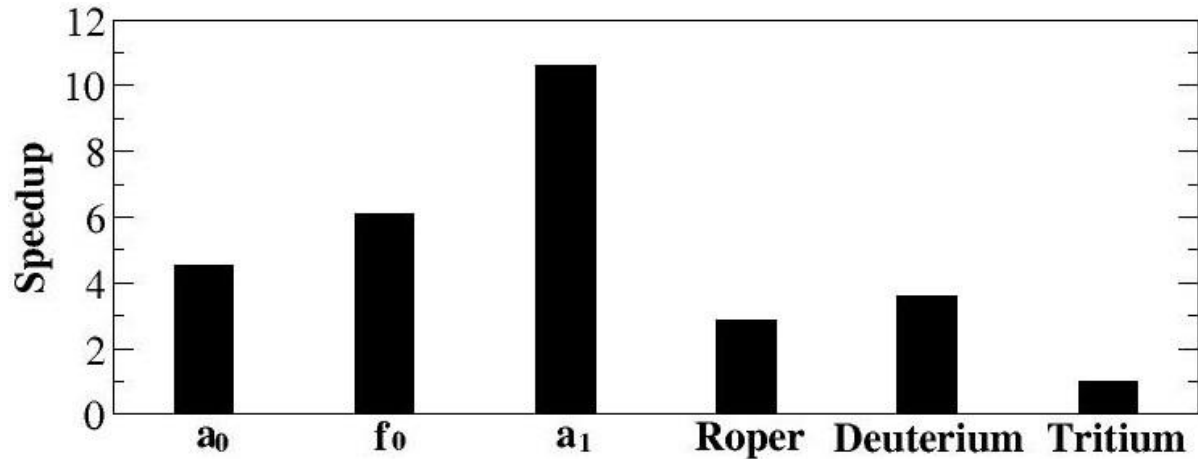
#	Name	Type	# graphs	$N_{\text{ref}}$	Complexity
1	$a_0$	$M \times M$	19041	256	$O(N^3)$
2	$f_0$	$M \times M \times M$	29600	256	$O(N^3)$
3	$a_1$	$M \times M \times M$	403072	128	$O(N^3)$
4	Roper	$B \times M$	84894	64	$O(N^4)$
5	Deuterium	$B \times B$	119191	64	$O(N^4)$
6	Tritium	$B \times B \times B$	6208	32	$O(N^5)$

- Testing platforms:

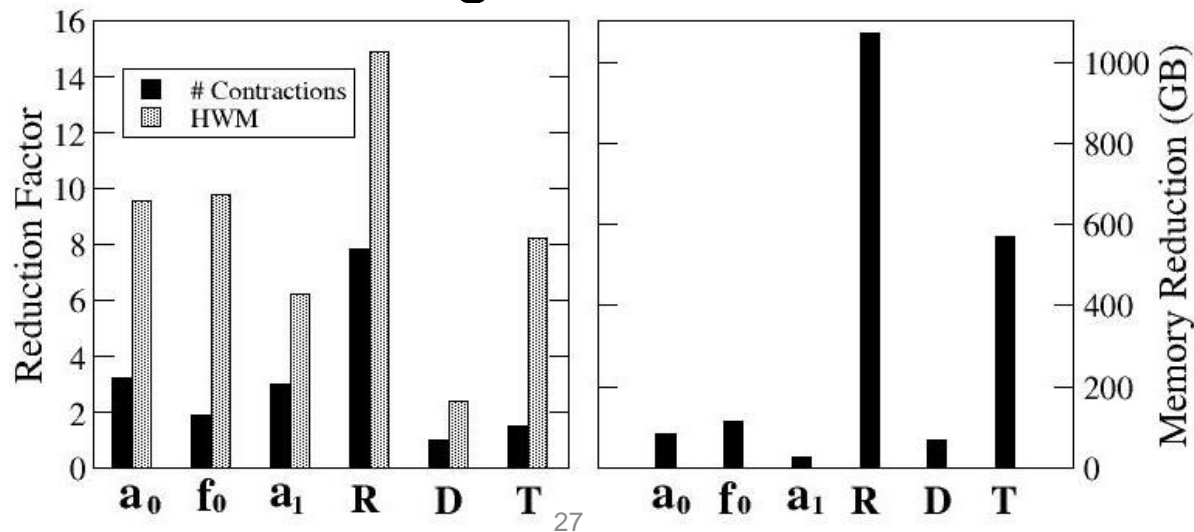
- Intel **KNL**: 64 cores, 192 GB memory
- **A100**: 4 A100 GPUs with 40 GB memory and NVlink. 1 AMD EPYC CPU (64 cores and 256 DDR4 memory).
- **MI100**: 8 MI100 GPUs with 32 GB memory and Infinity Fabric. 2 AMD EPYC 7502 CPU each has 32 cores. 1 TB host DDR5 memory.
- **MI250**: 4 MI250X with 8 separate MI250 GPUs each has 64 GB memory and Infinity Fabric. 1 AMD EPYC 7A53 CPU (64 cores and 512 DDR5 memory).

# Performance Results

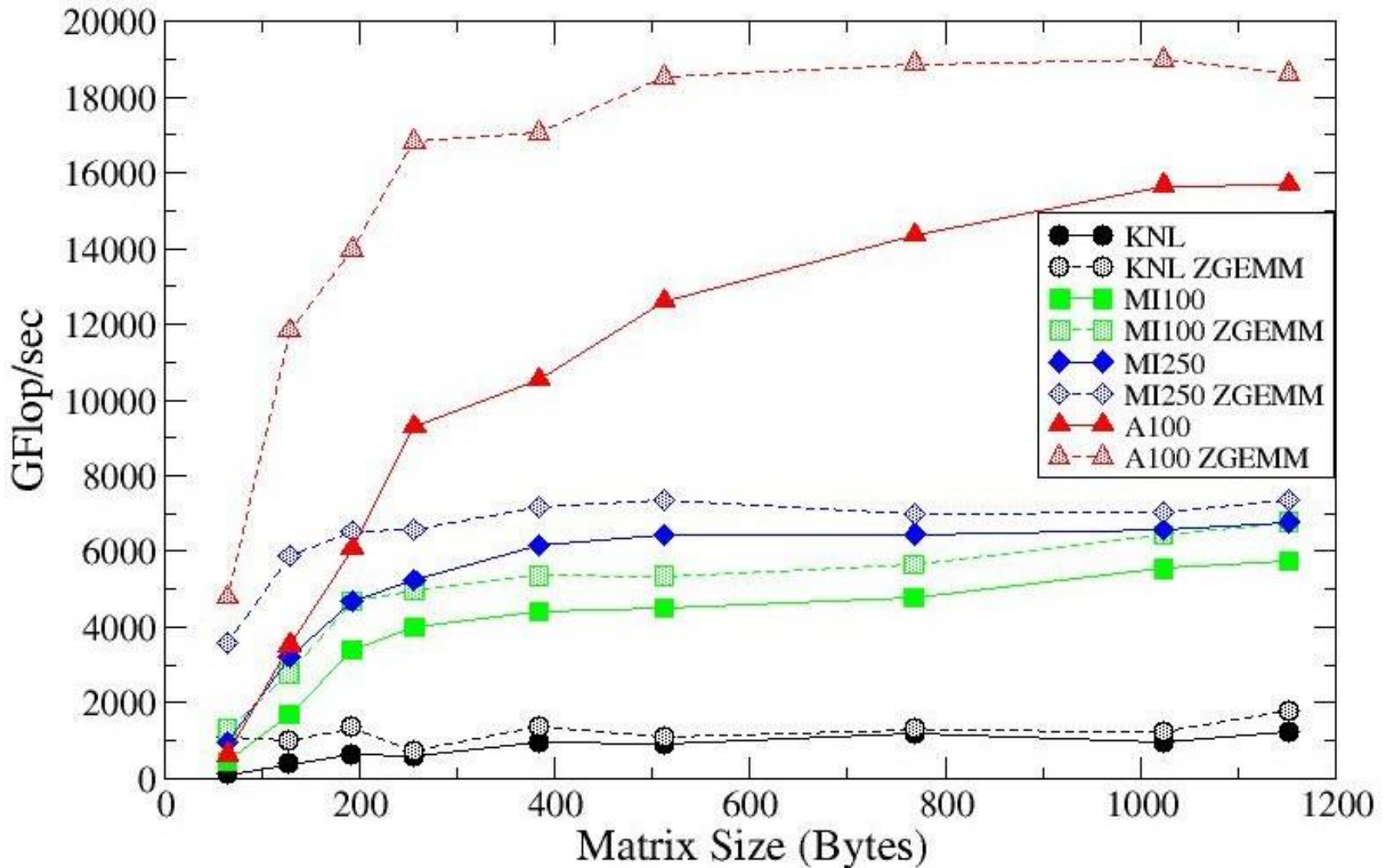
- Graph Classification



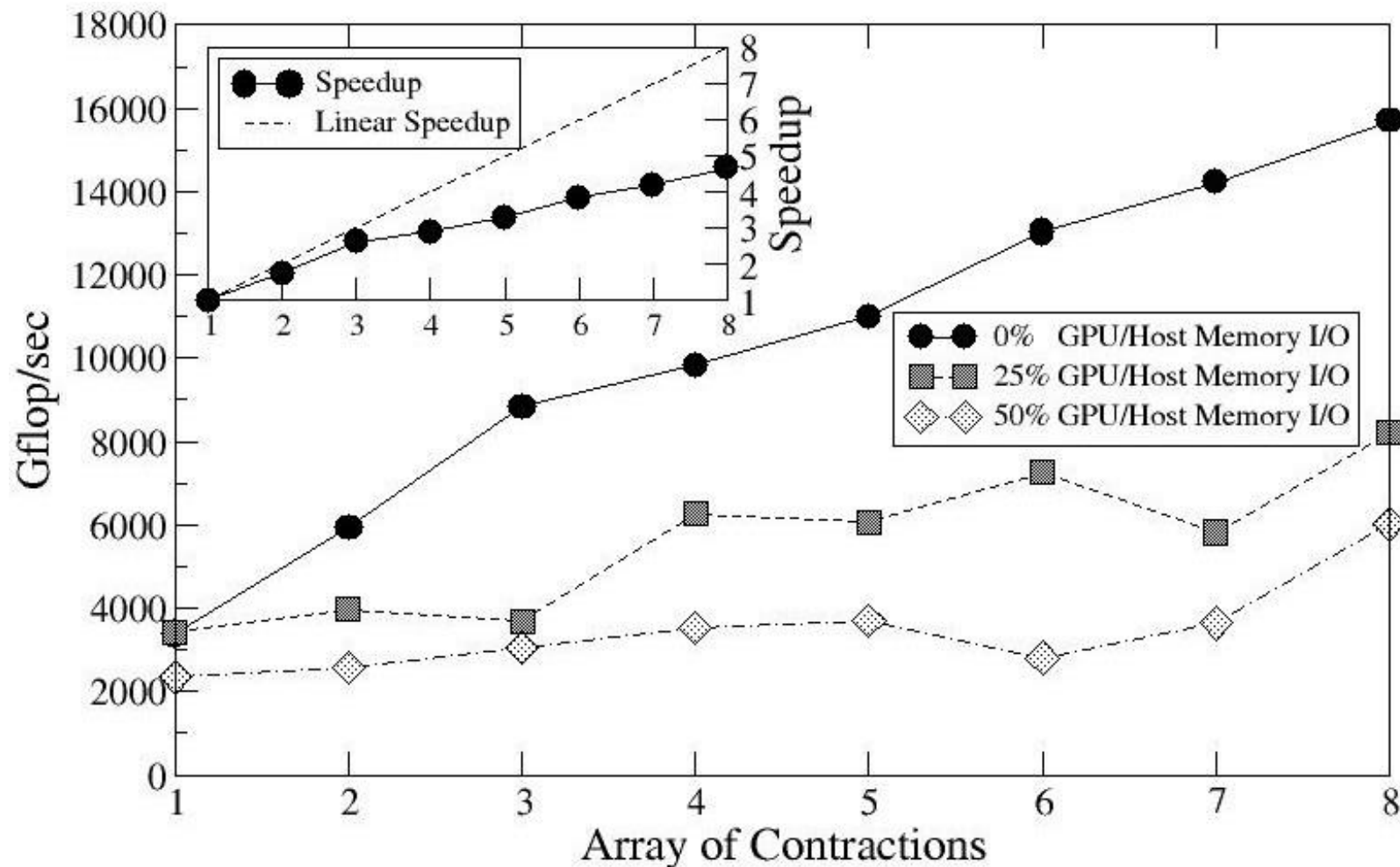
- Graph contraction algorithm



# Hadron Contraction vs Batched zgemms



# Hadron Contraction on Multiple GPUs

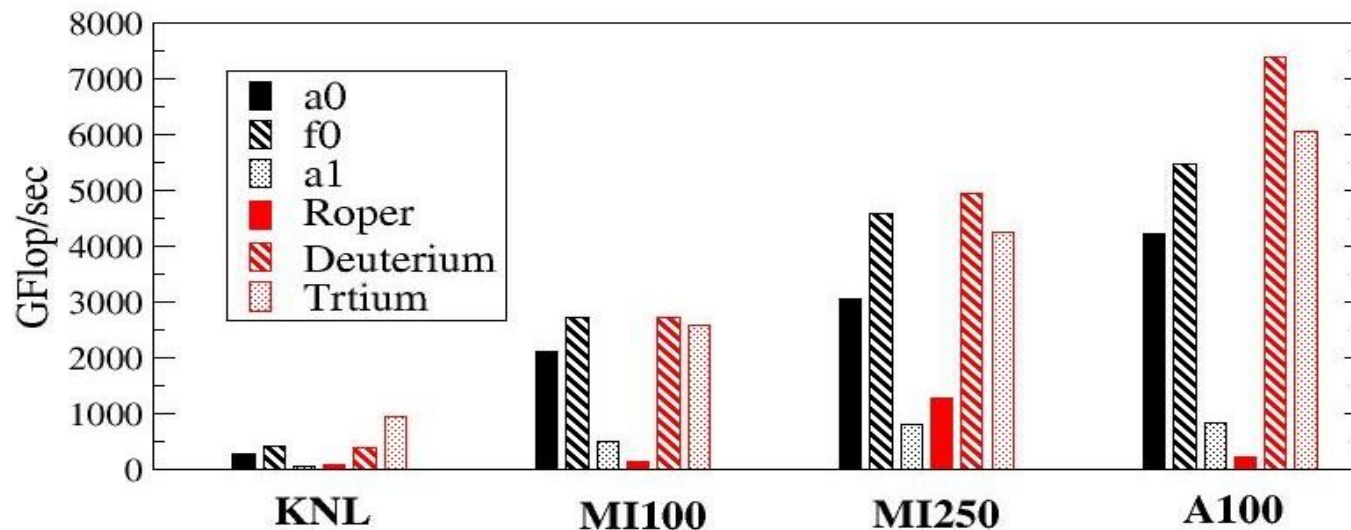


# Redstar Performance on a Single GPU and a KNL

- Timing information for one time slice (MI100)

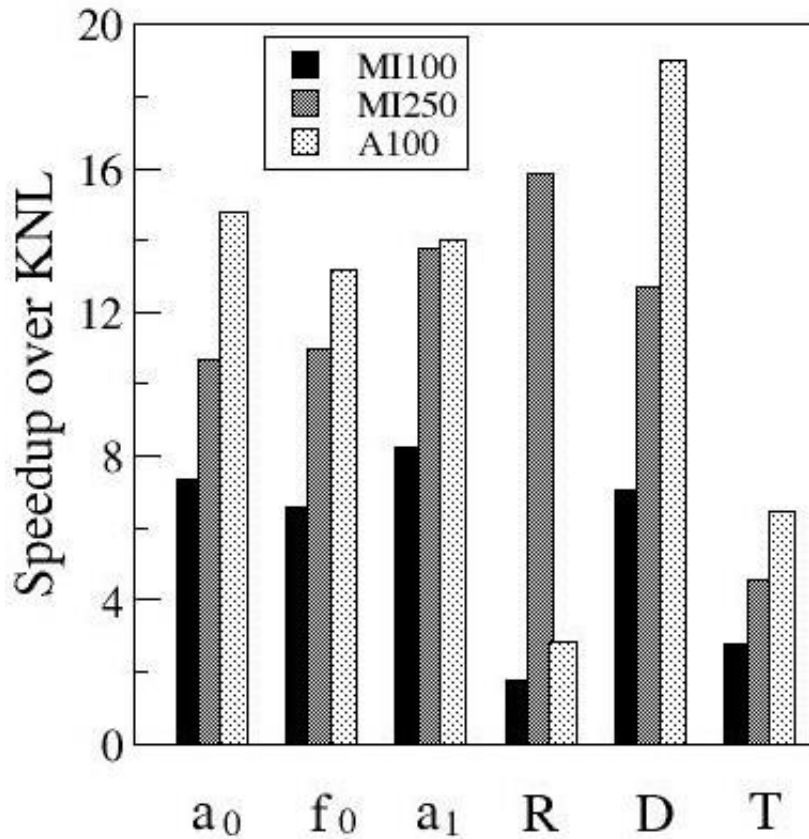
Name	Classification	DAG	Contraction	Time
$a_0$	12.5	19.1	10.5	42.1
$f_0$	28.5	31.9	16.6	77.0
$a_1$	925.5	929.4	117.1	1972
<i>Roper</i>	37.2	137.9	1924.1	2099.2
<i>Deuterium</i>	122.4	191.7	501.4	811.5
<i>Tritium</i>	37.1	19.3	2952.3	3008.7

- Performance on a GPU and a single KNL.

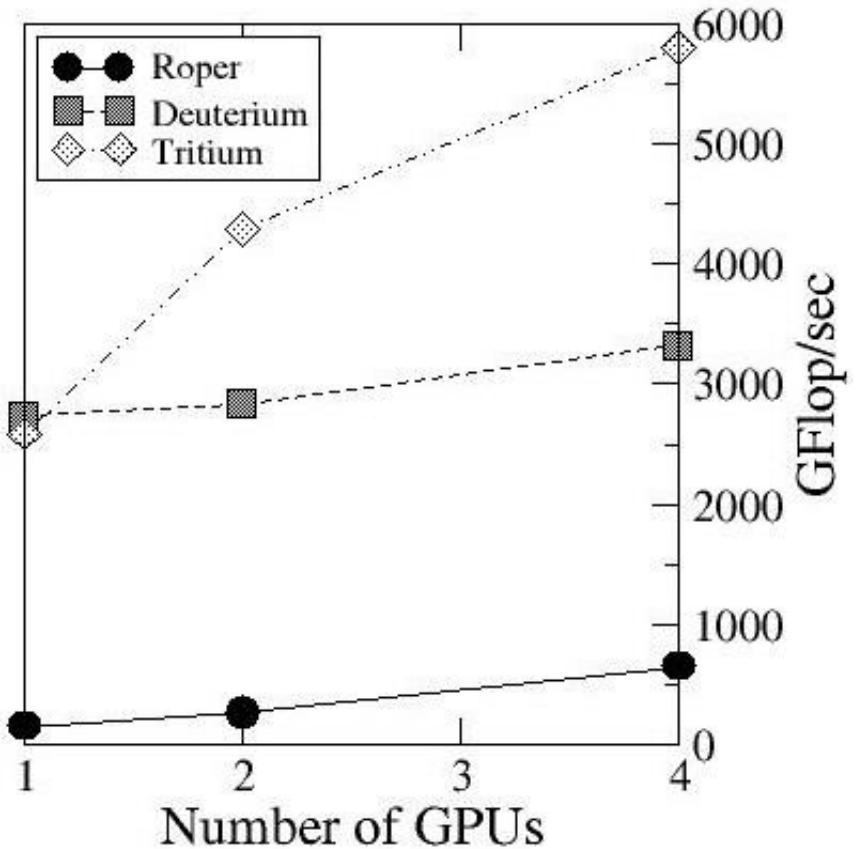


# Redstar Performance

## Speedup for one GPU over one KNL



## Multi-GPU performance MI100



# Conclusion and Future Work

## Contributions:

- *Redstar* converts symbolic calculations (*Wick contractions*) into graphs.
- *Redstar* utilizes canonical labeling to identify the unique graphs efficiently.
- *Redstar* utilizes graph algorithms to find optimal contractions path (calculation orders) for all graphs (*Wick contractions* terms) to reduce the overall computing cost and memory footprint.
- *Redstar* performs very well calculating correlation functions especially on GPUs.
- *Redstar* is an end-to-end production software suite.
  - (Frontier, Perlmutter...)

## Future Work:

- Apply graph partition technique on the **execution DAG** to further reduce data movements among GPUs.
- Enable *Redstar* to run on Intel GPUs and multiple hosts.



---

*Thank you !*

**Questions?**