Simulation Model of an HPC System for Super Charm-Tau Factory

D. WIENS, I. CHERNYKH, I. LOGASHENKO, F. KOLPAKOV, V. VOROBIEV

BINP SB RAS, FRC ICT, SSCC, ICMMG SB RAS, Novosibirsk

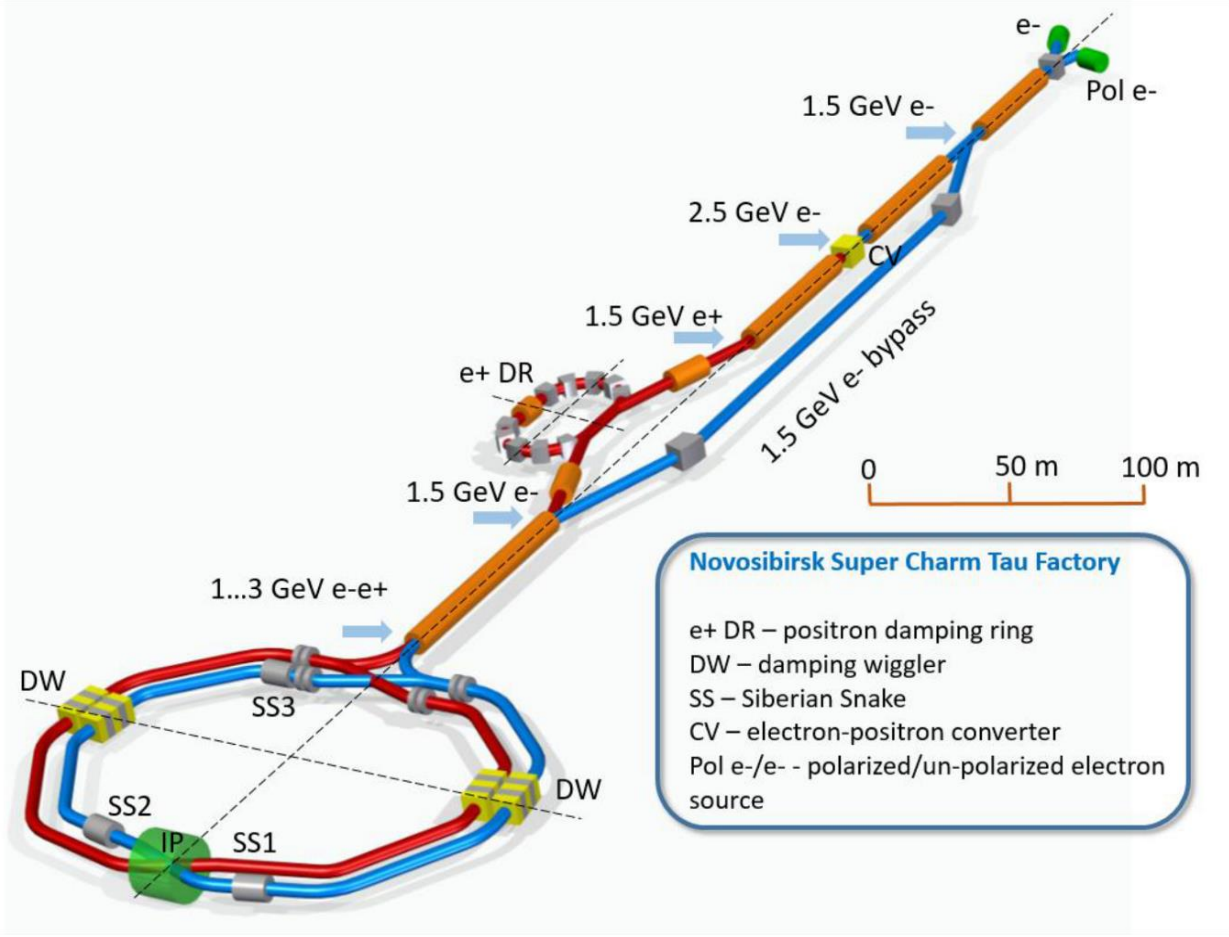
This work was supported by the Russian Science Foundation (Grant No. 19-72-20114)

Super Charm-Tau Factory Project

The «Super Charm-Tau Factory» project, which is a symmetric electron-positron collider of ultrahigh luminosity with a beam energy at the mass center system from 2 to 6 GeV, is developed at the BINP SB RAS.

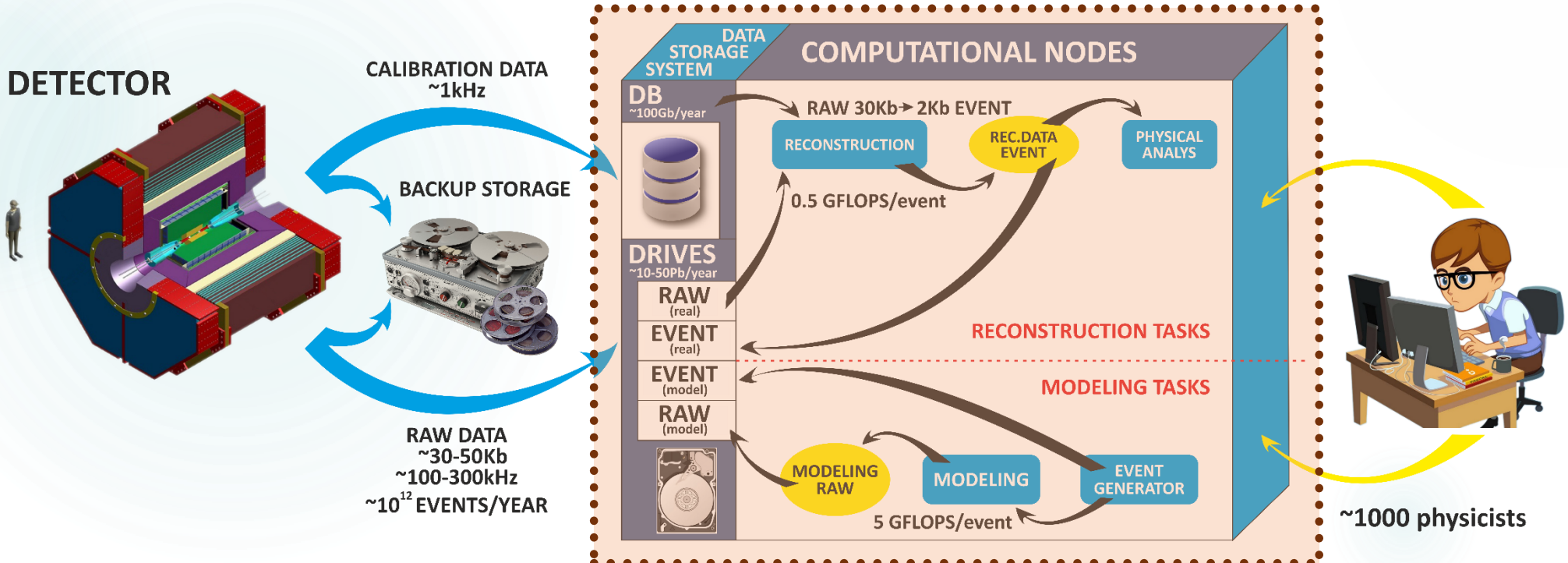
This project comprises a unique accelerating-storage complex with a luminosity of 10^{35} $cm^{-2}s^{-1}$ and a universal elementary particle detector.

The main goal of experiments carried out on the SCTF is to study the properties of tau lepton and charmed particles, subject the existing microworld theory and Standard Model to high-presicion verification, and to search for phenomena not described within the framework of this theory.



Super Charm-Tau Factory Project

In the course of the experiments, about 100 petabytes of "RAW" data is accumulated from the elementary particle detector of the SCTF. An important role in the project is played by the system for data processing and storage, whose tasks include the primary data processing, data transfer to long-term data storage system (decades), data extraction from the storage system for processing and processing using high-tech computing (HPC) systems. Specialized software should allow one to analyze the accumulated data by a collective of about 1000 physicists. The development of the data analysis algorithms and the optimization of the detector structure are carried out using modeling data generated via software for modeling of experiments.



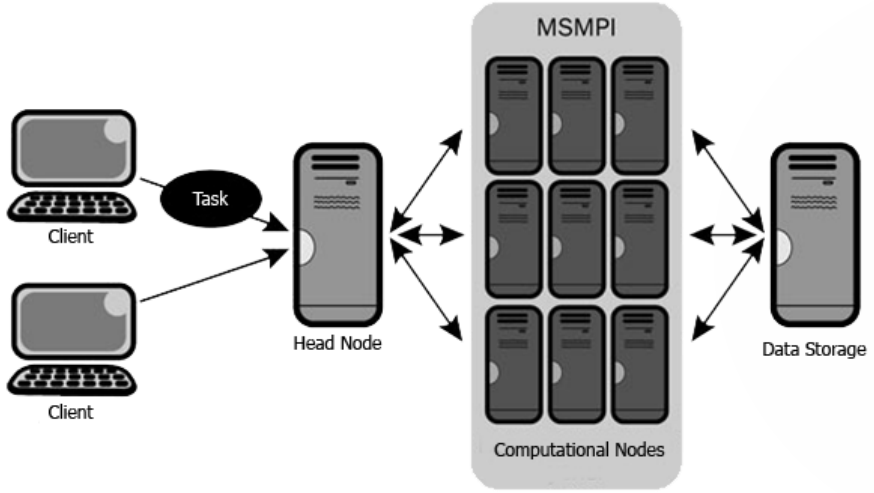
General knowledge of HTC Systems

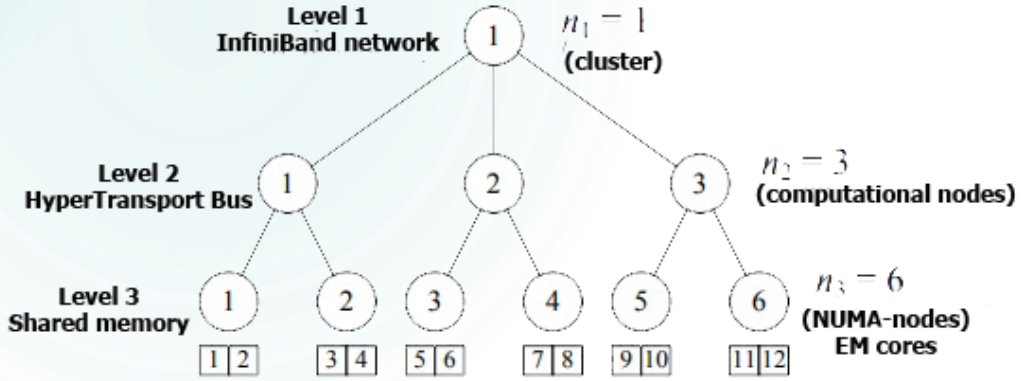
Main purpose is calculating the parameters of the computing system necessary for processing and storing the results of operation of the Super Charm-Tau factory after its commissioning.

Cluster is a composition of computers, communication network between them and software for parallel information processing.

Tasks:

- Job flow maintenance;
- Resource allocation;
- Balancing resources;
- Providing fault tolerance;
- Ensuring energy efficiency.





Hierarchical organization of the communication environment

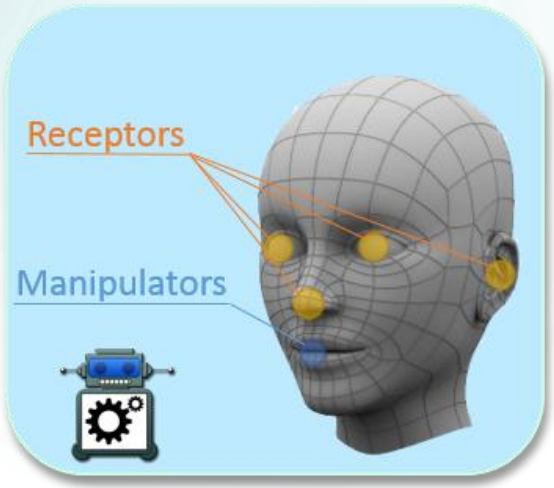
	Node 1	Node 2	…	Node N
Node 1	0	1	…	4
Node 2	1	0	…	2
…	…	…	0	…
Node N	4	2	…	0

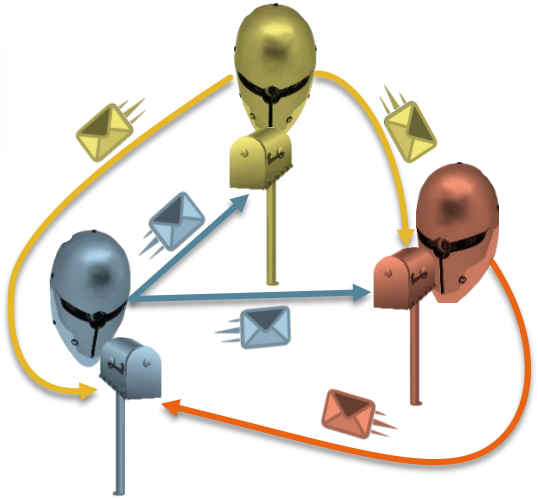
Levels of closest common ancestors matrix

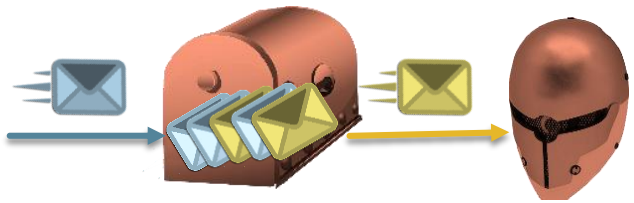
Multiagent approach

Agent modeling is an advanced approach to modeling systems containing autonomous and interacting intelligent agents.

An intelligent agent is an entity existing in a medium and possessing sensors for perceiving this medium and executive mechanisms for affecting this medium. The agents operate asynchronously according to their laws and interact with other agents in order to reach common goals. During operation, a software agent can change both the outer medium and its own behavior. A distributed, asynchronous behavior is quite important for constructing a simulation model of a supercomputer as it is virtually impossible to ensure the centralized control of tens and hundreds of millions of cores.







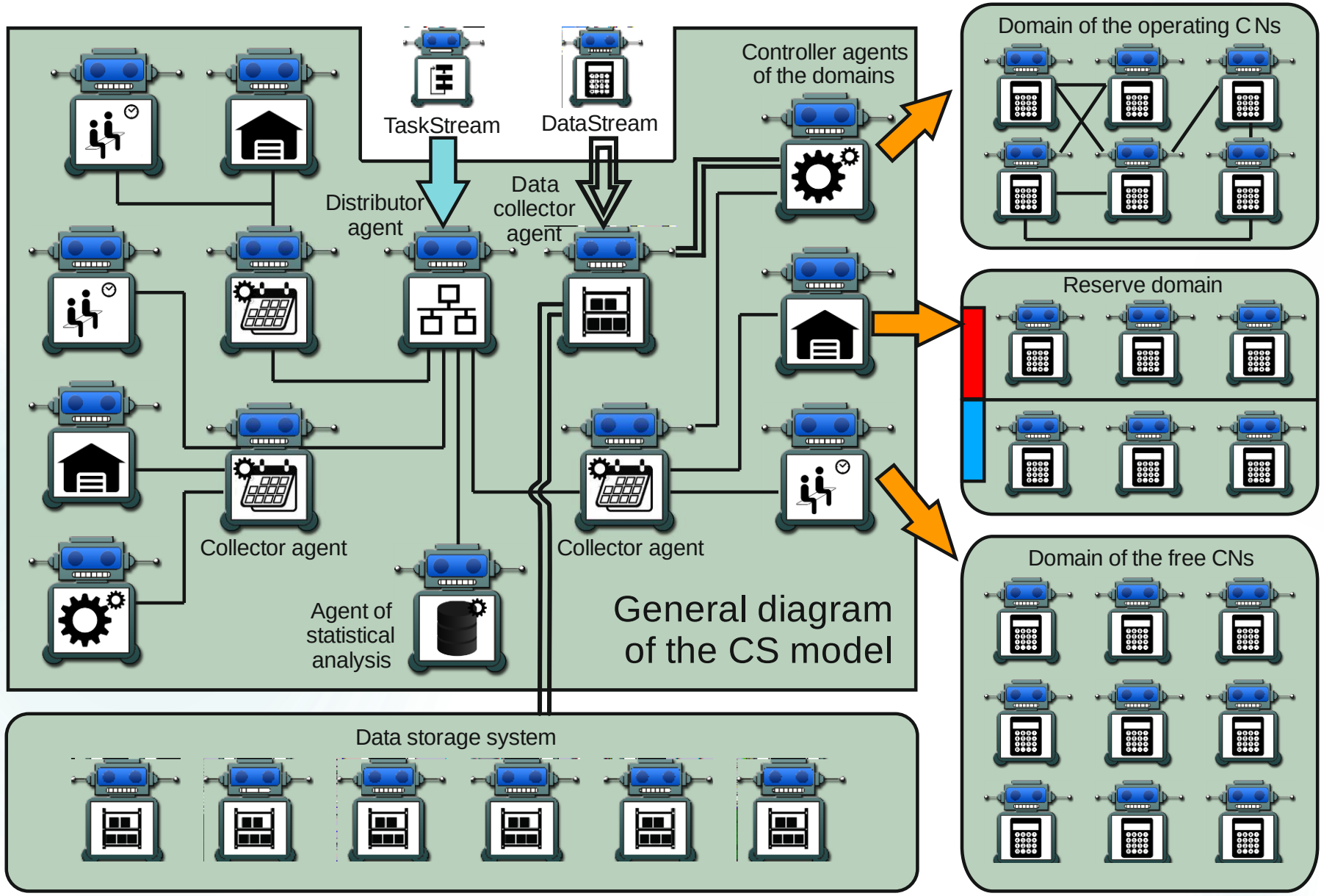
AGNES multiagent system

The AGNES system, initially developed (ICMMG SB RAS) for telecommunication and information networks, also demonstrated its efficiency in modeling the operation of a distributed control system and executing high-performance parallel programs.

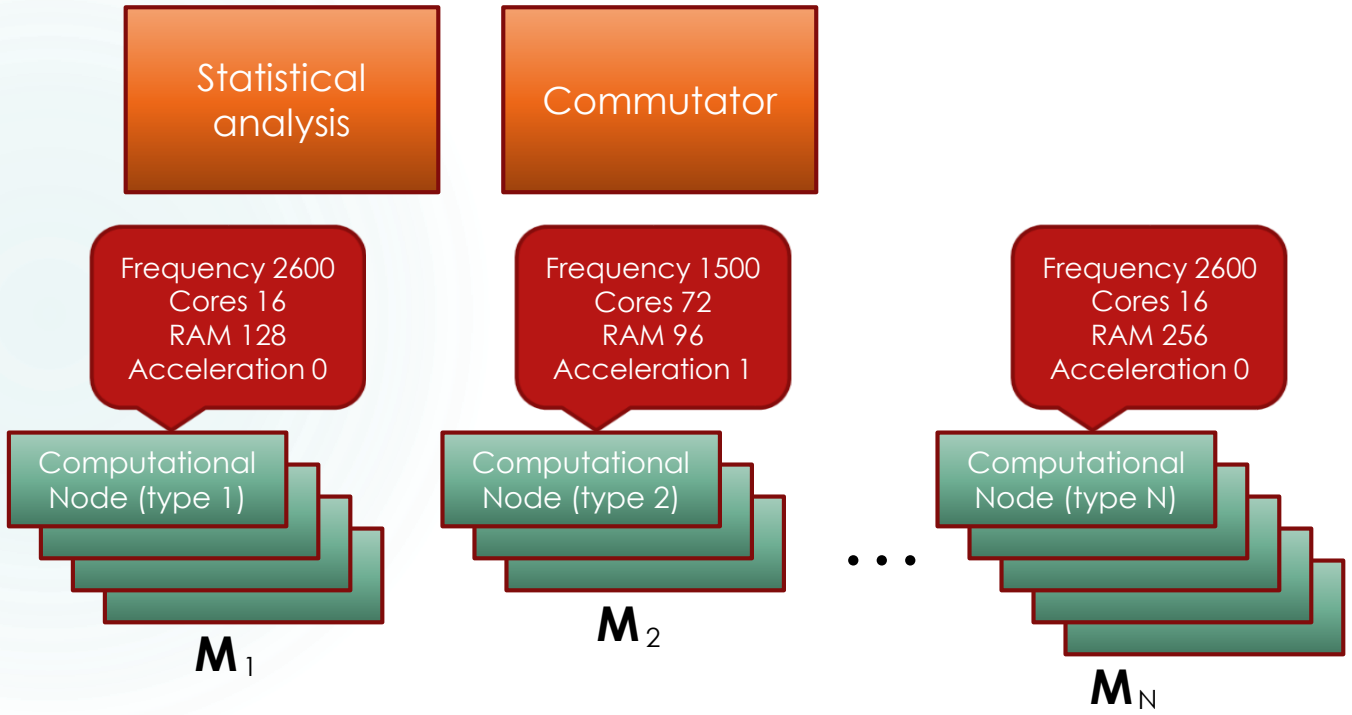
The AGNES package is based on Java Agent Development Framework (JADE), serving as a means for creating Java based MAS. JADE is a powerful tool for developing Java based multiagent systems. It is important for modeling large computations that JADE is a FIPA-compatible, distributed agent platform, which can use one or several computers (network nodes), each having only one operationing JAVA machine.

All interaction between JADE agents occurs by exchanging messages according to FIPA specification. The key property of a JADE agent is a set of its «behaviors». An agent's lifecycle ends when this agent has no active behaviors. AGNES uses the advantages provided by JADE and expands the multiagent system to a modeling system.

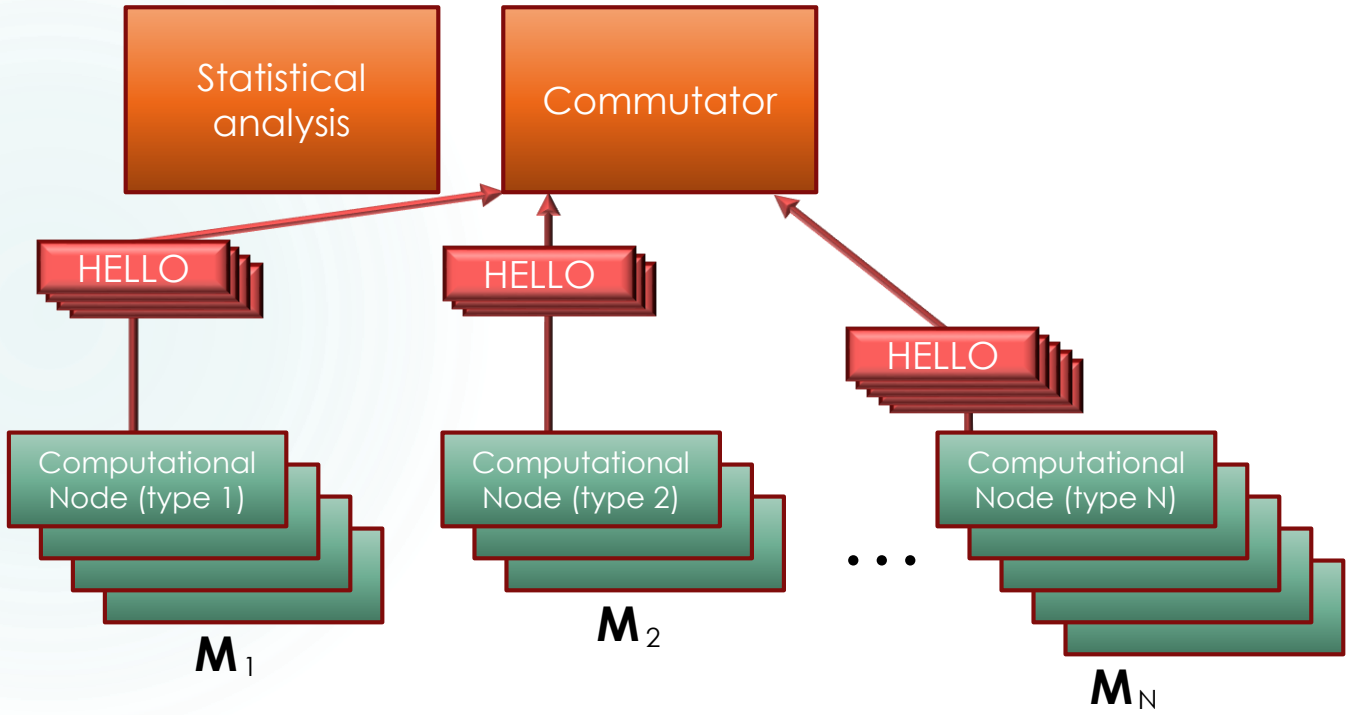
Model of HTC System. General scheme



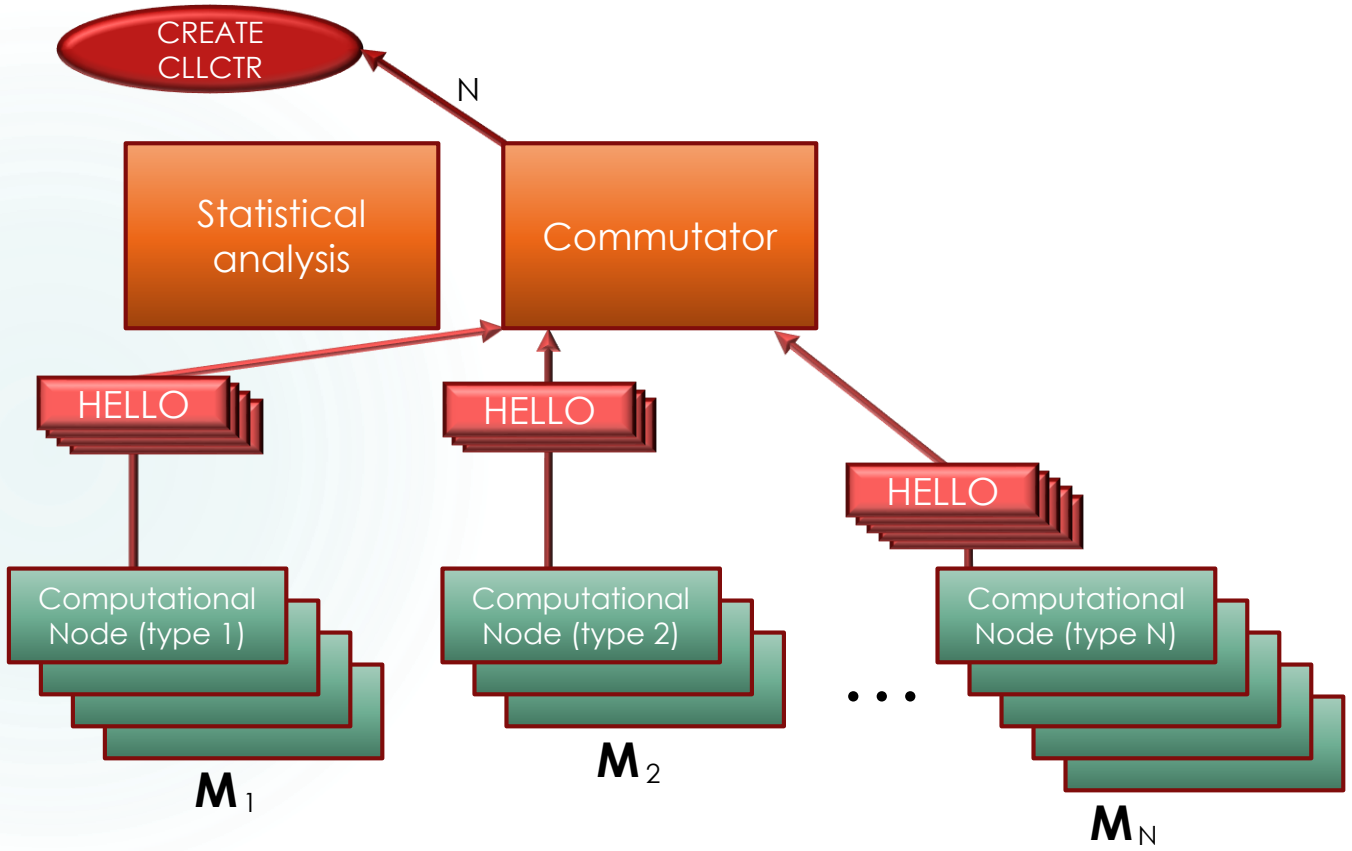
Model of HTC System. Initialization.



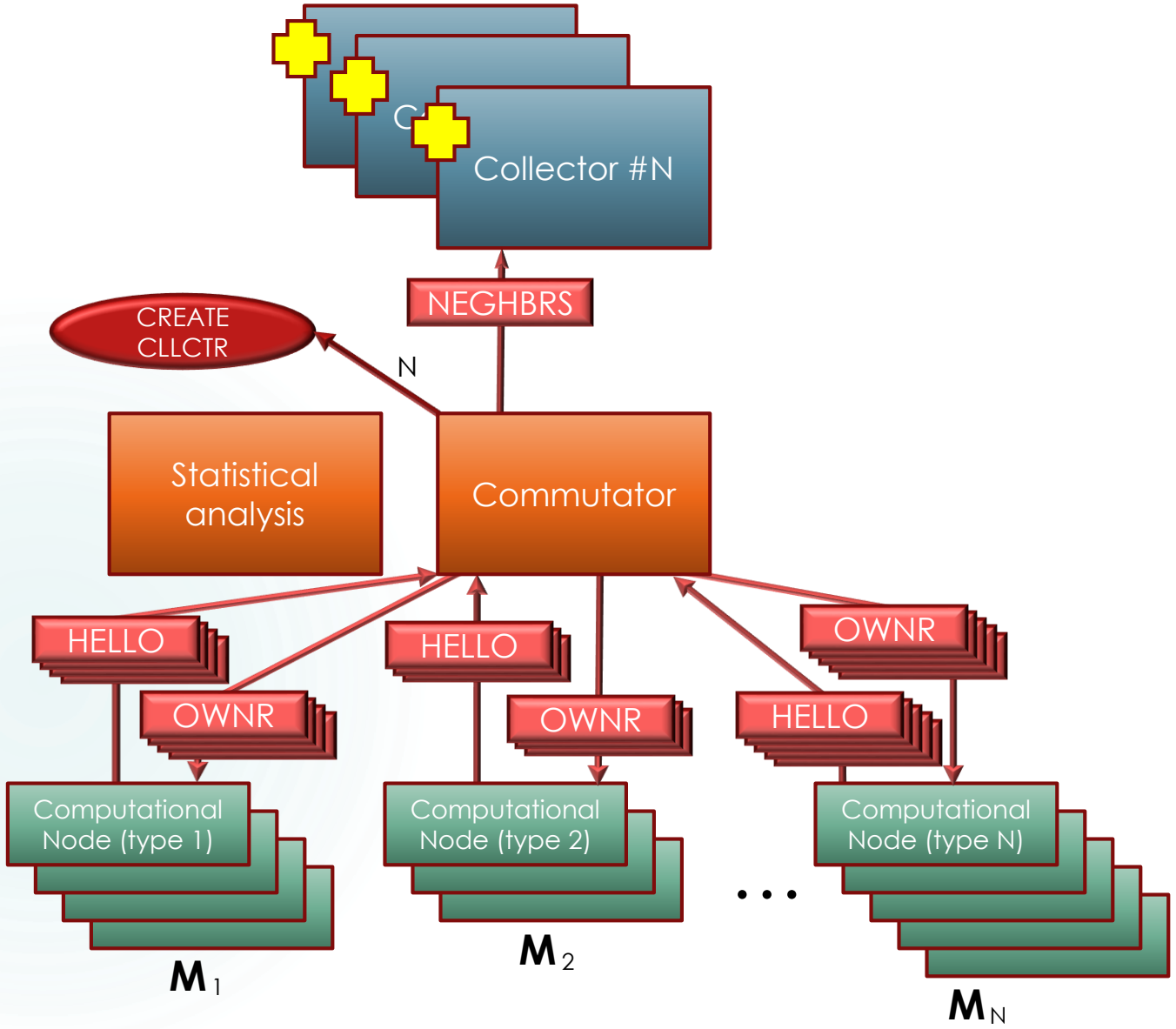
Model of HTC System. Initialization.



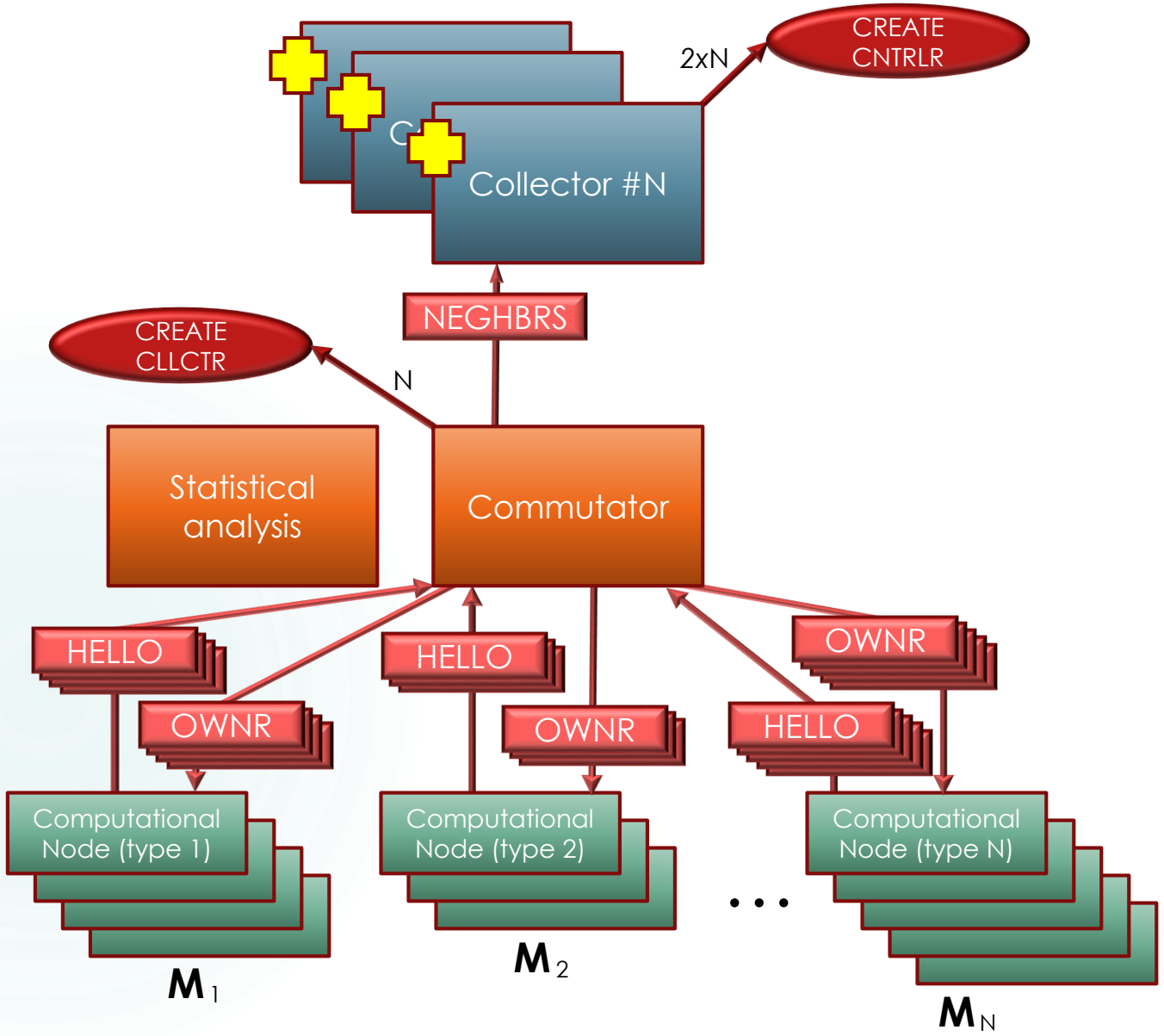
Model of HTC System. Initialization.



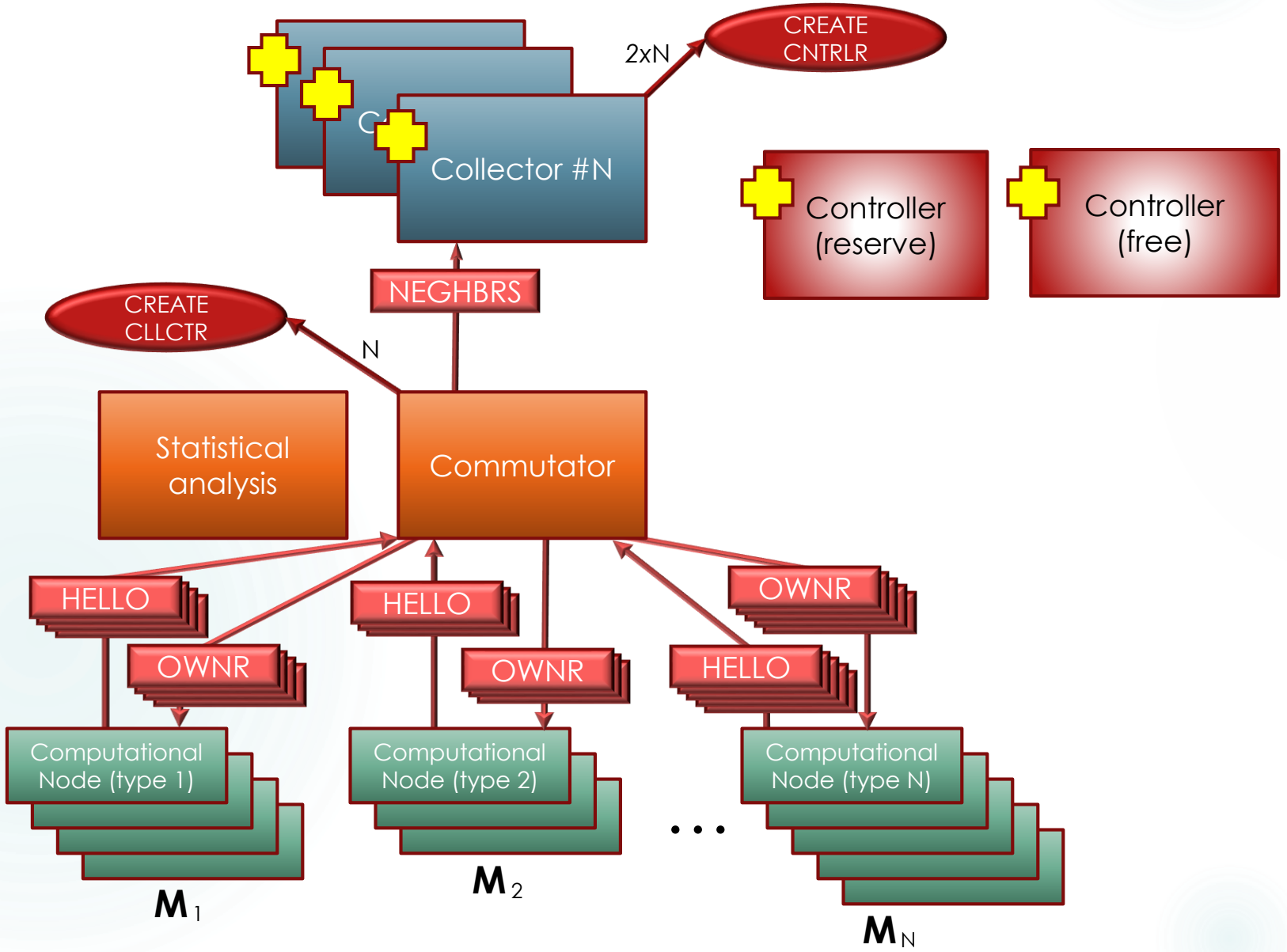
Model of HTC System. Initialization.



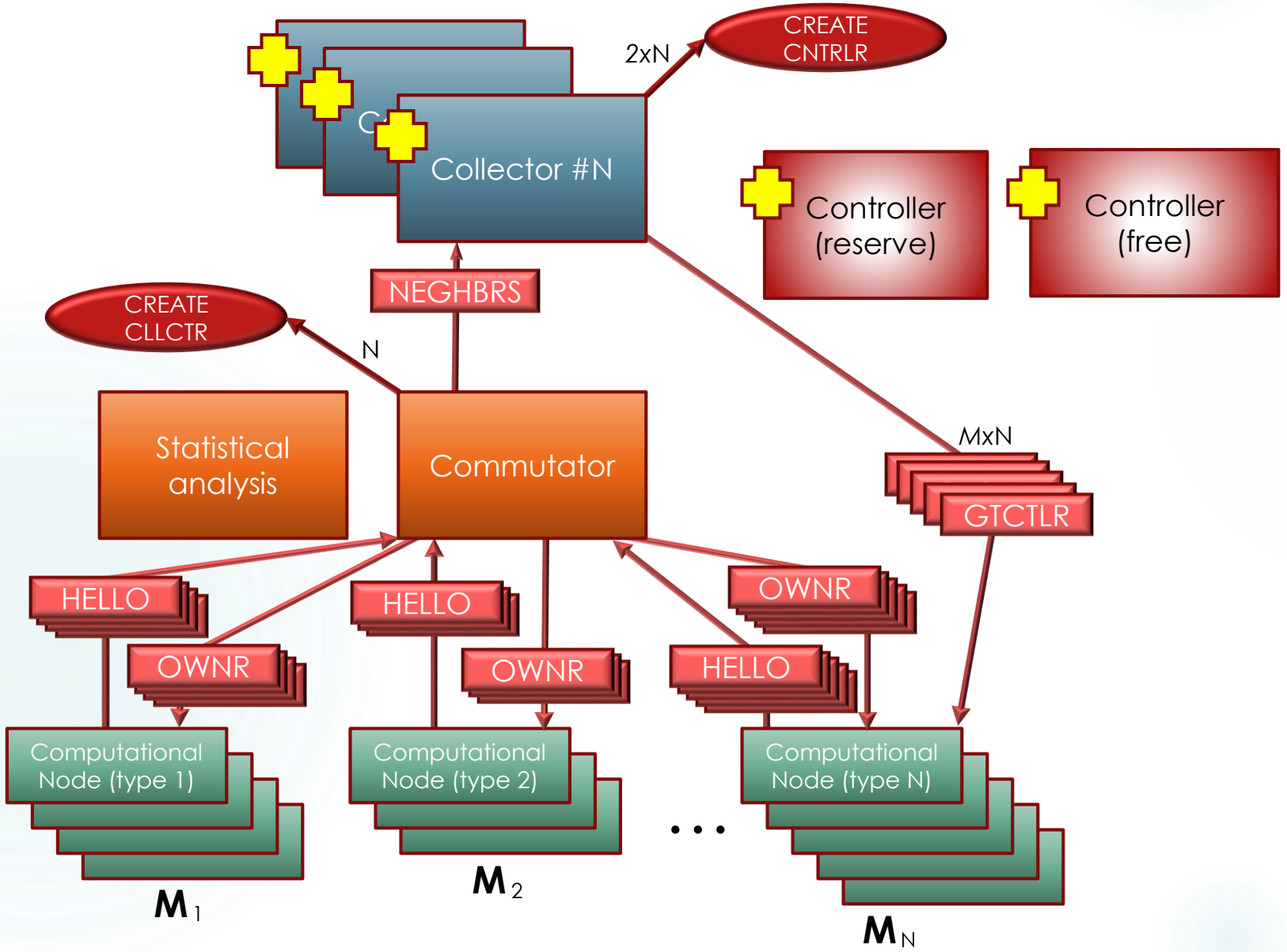
Model of HTC System. Initialization.



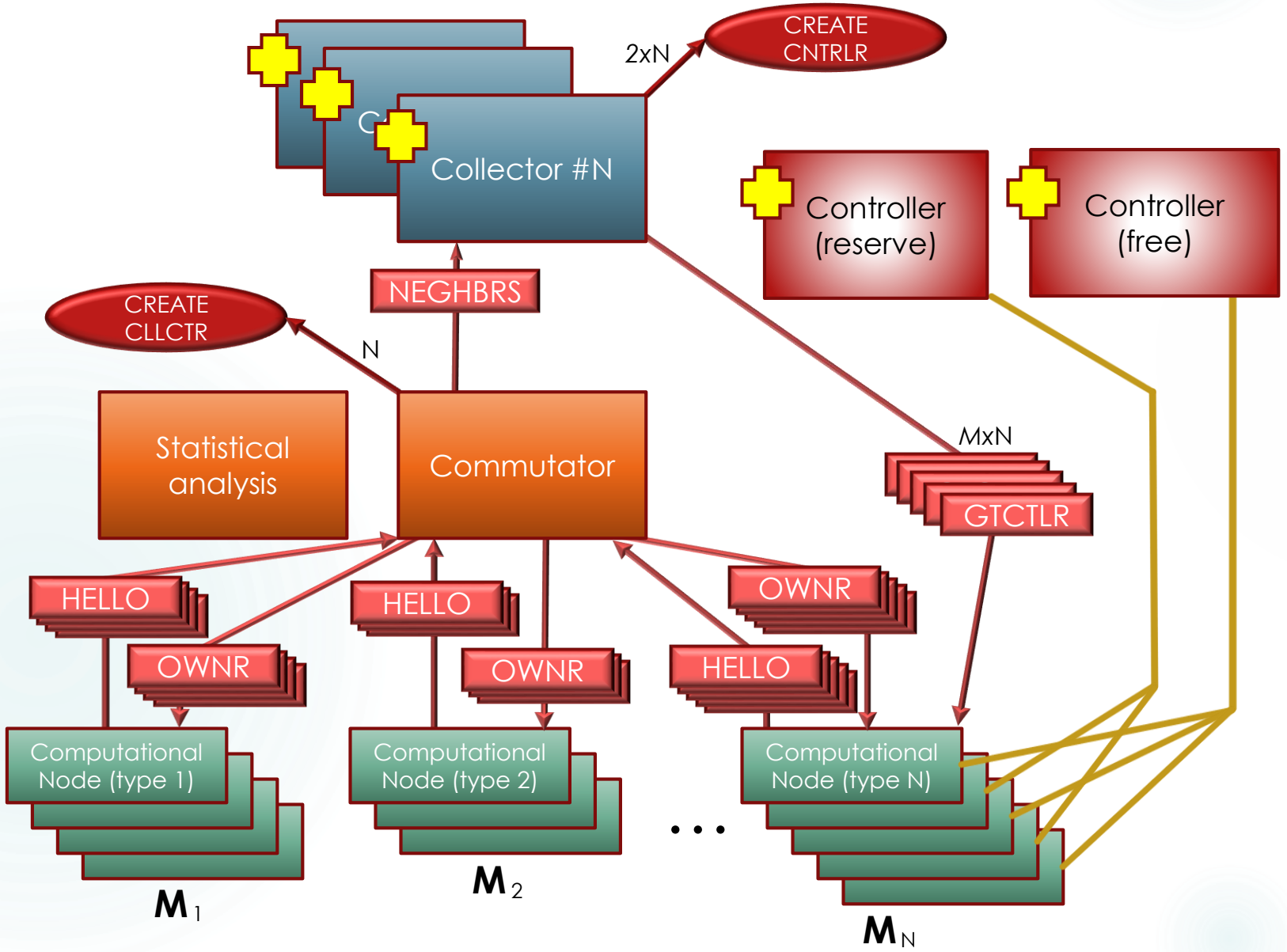
Model of HTC System. Initialization.



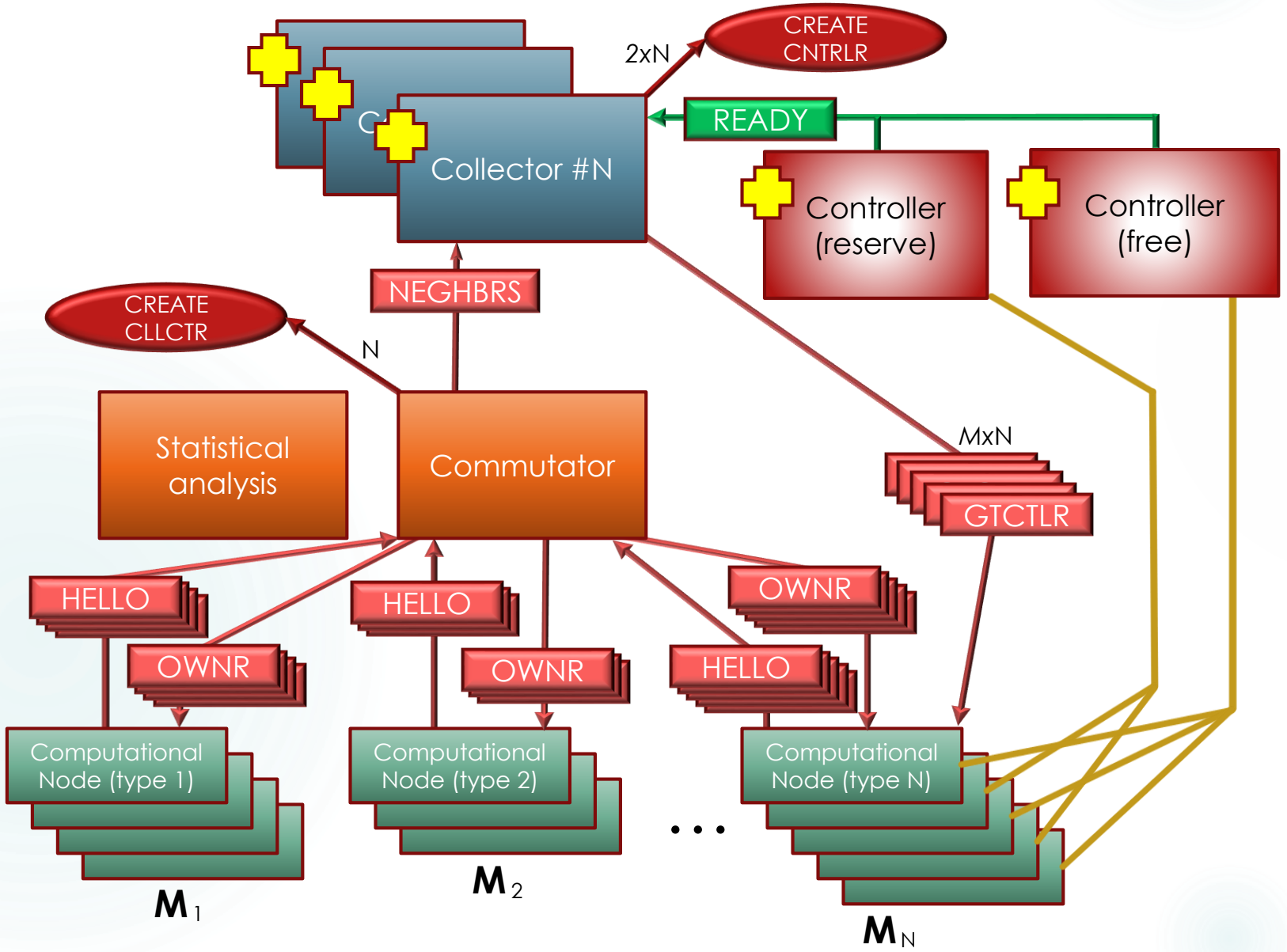
Model of HTC System. Initialization.



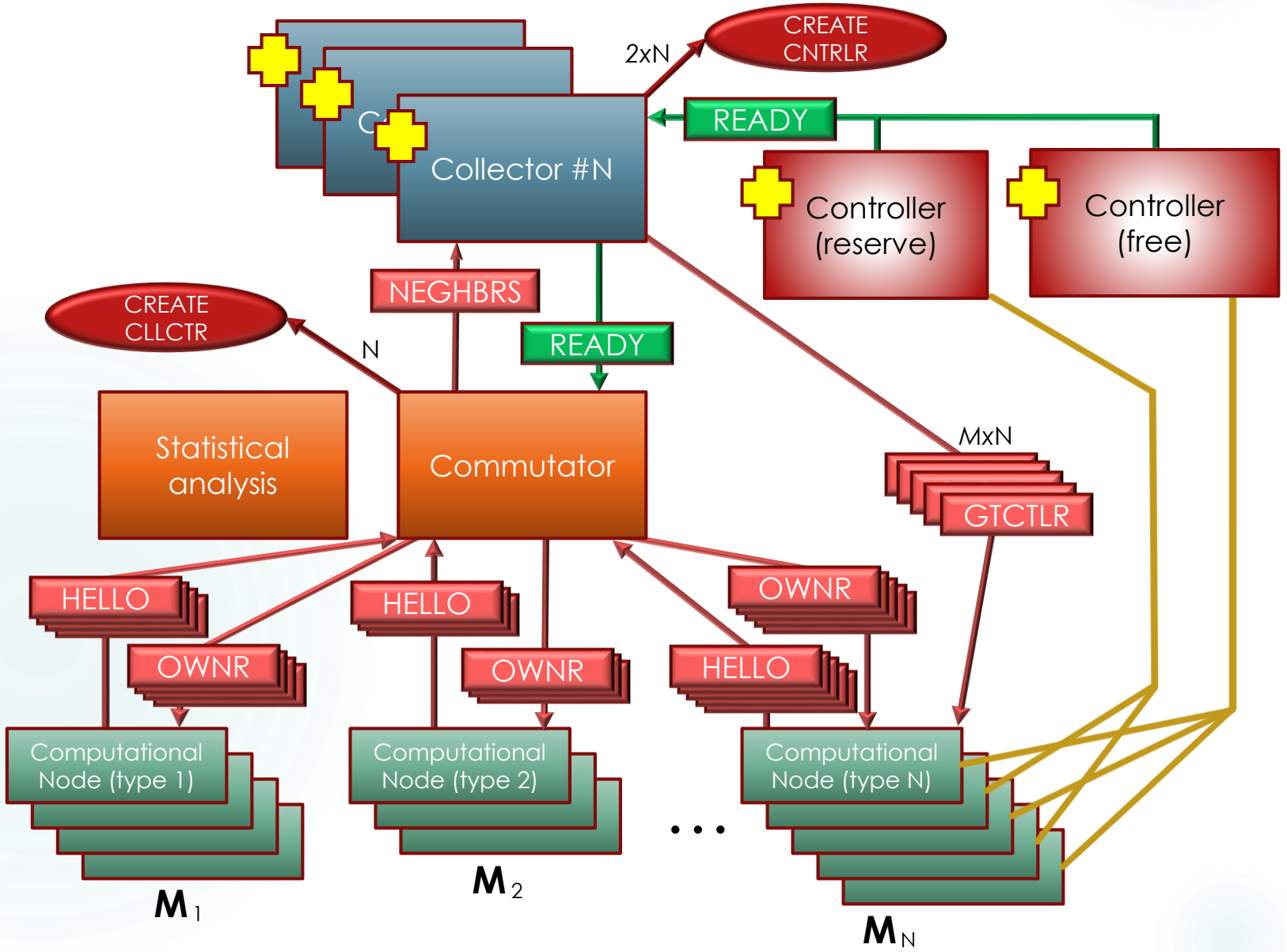
Model of HTC System. Initialization.



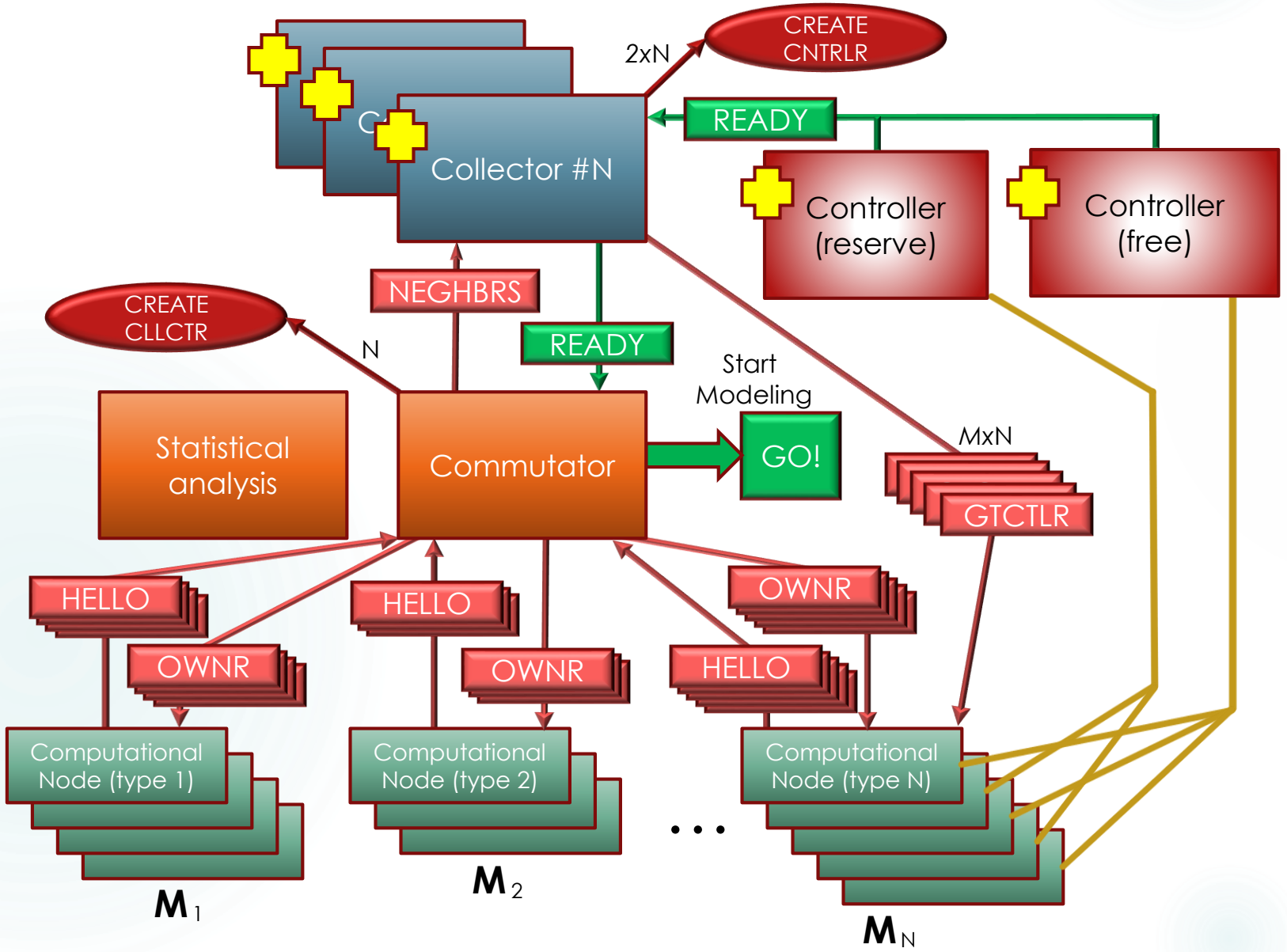
Model of HTC System. Initialization.



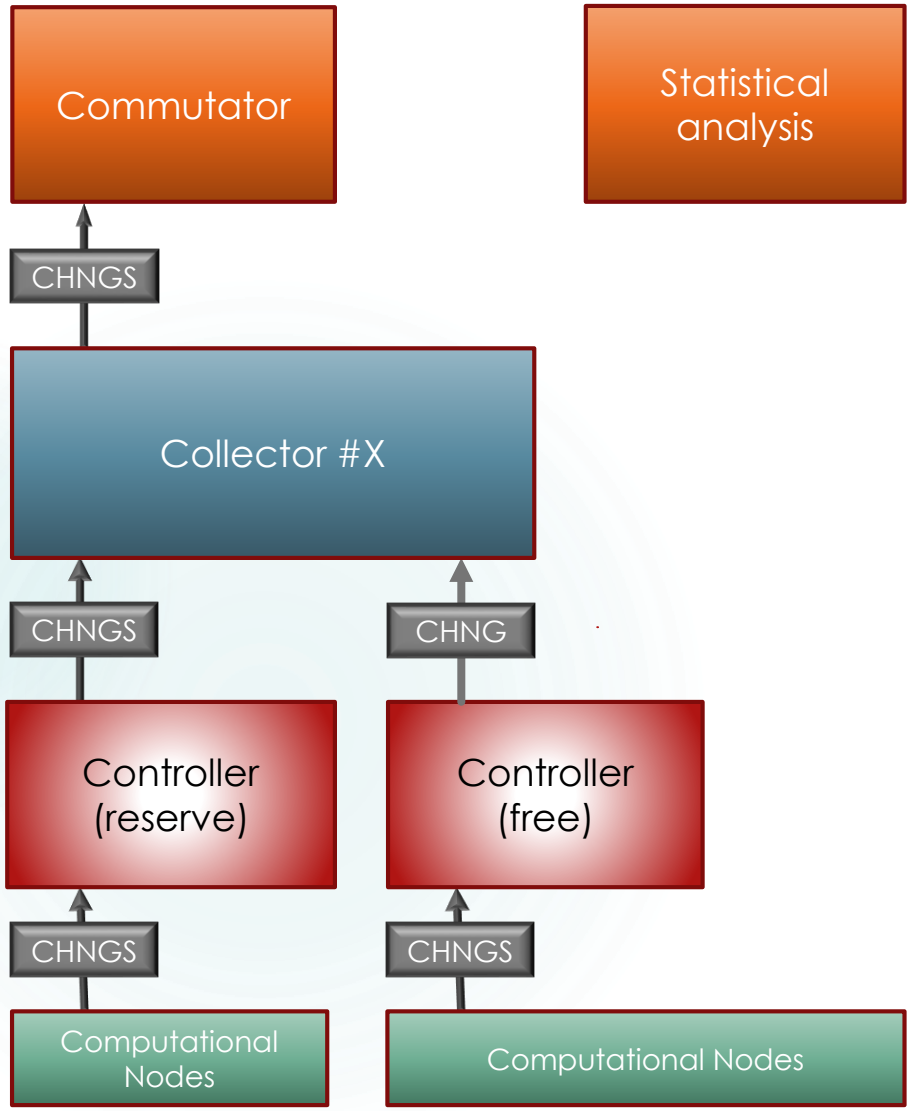
Model of HTC System. Initialization.



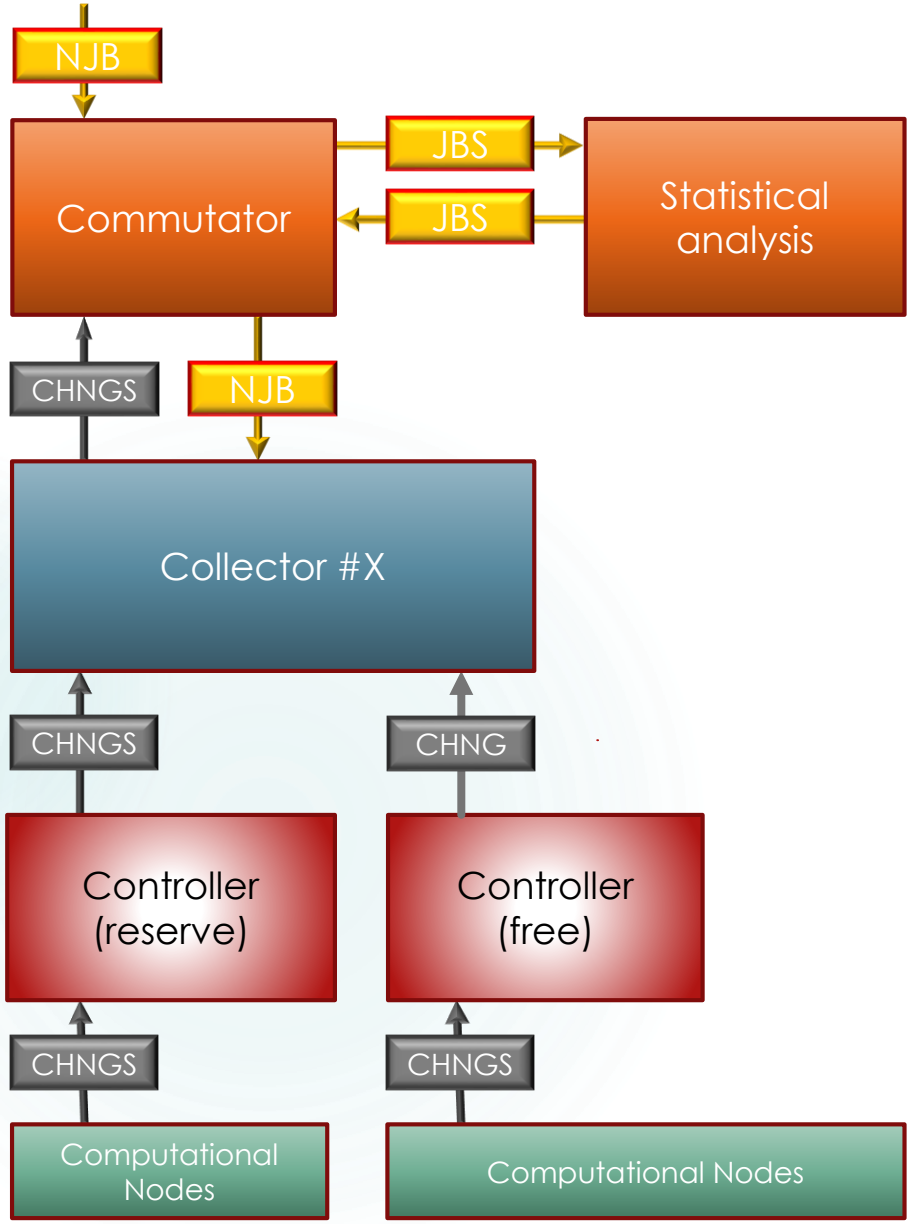
Model of HTC System. Initialization.



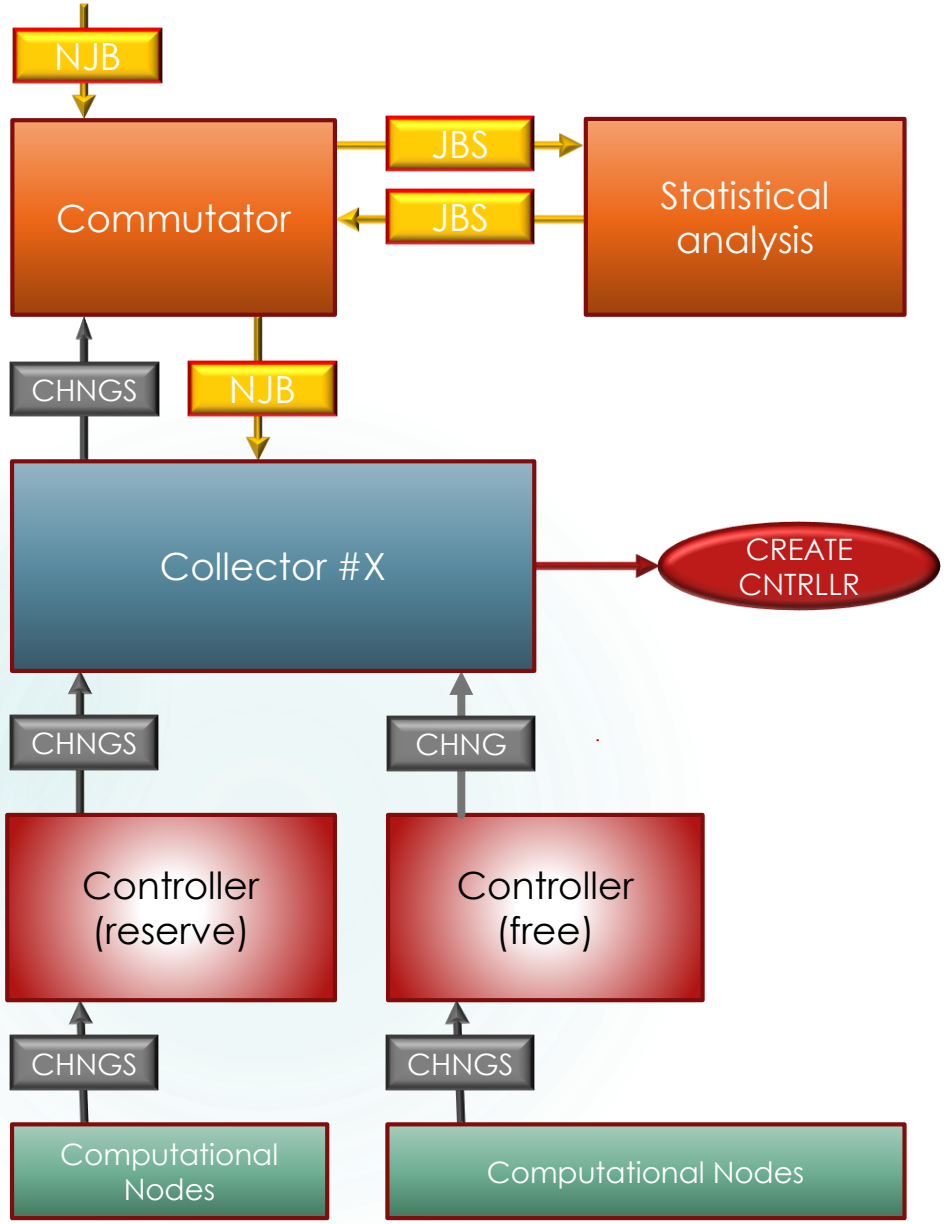
Model of HTC System. Task processing



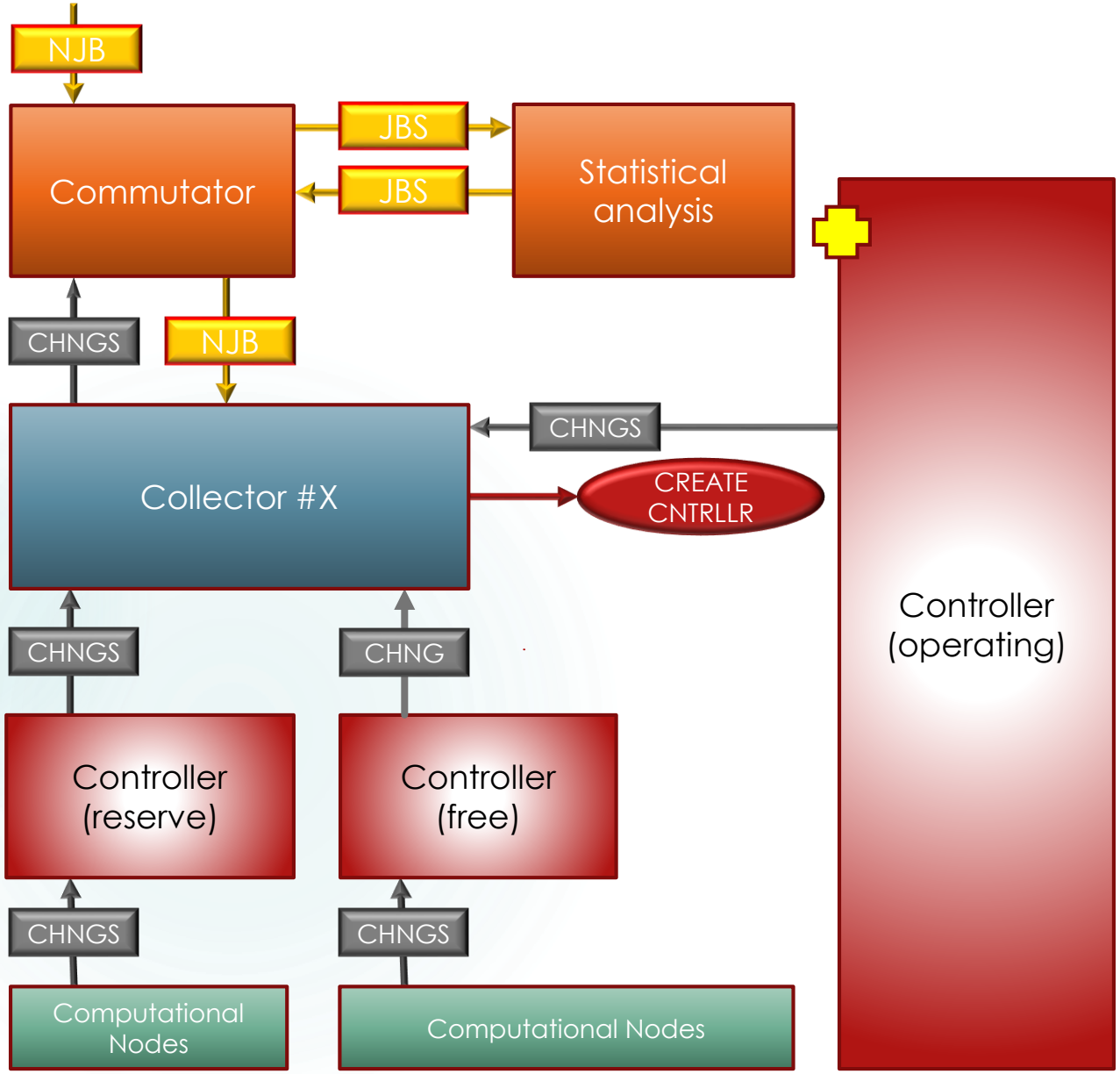
Model of HTC System. Task processing



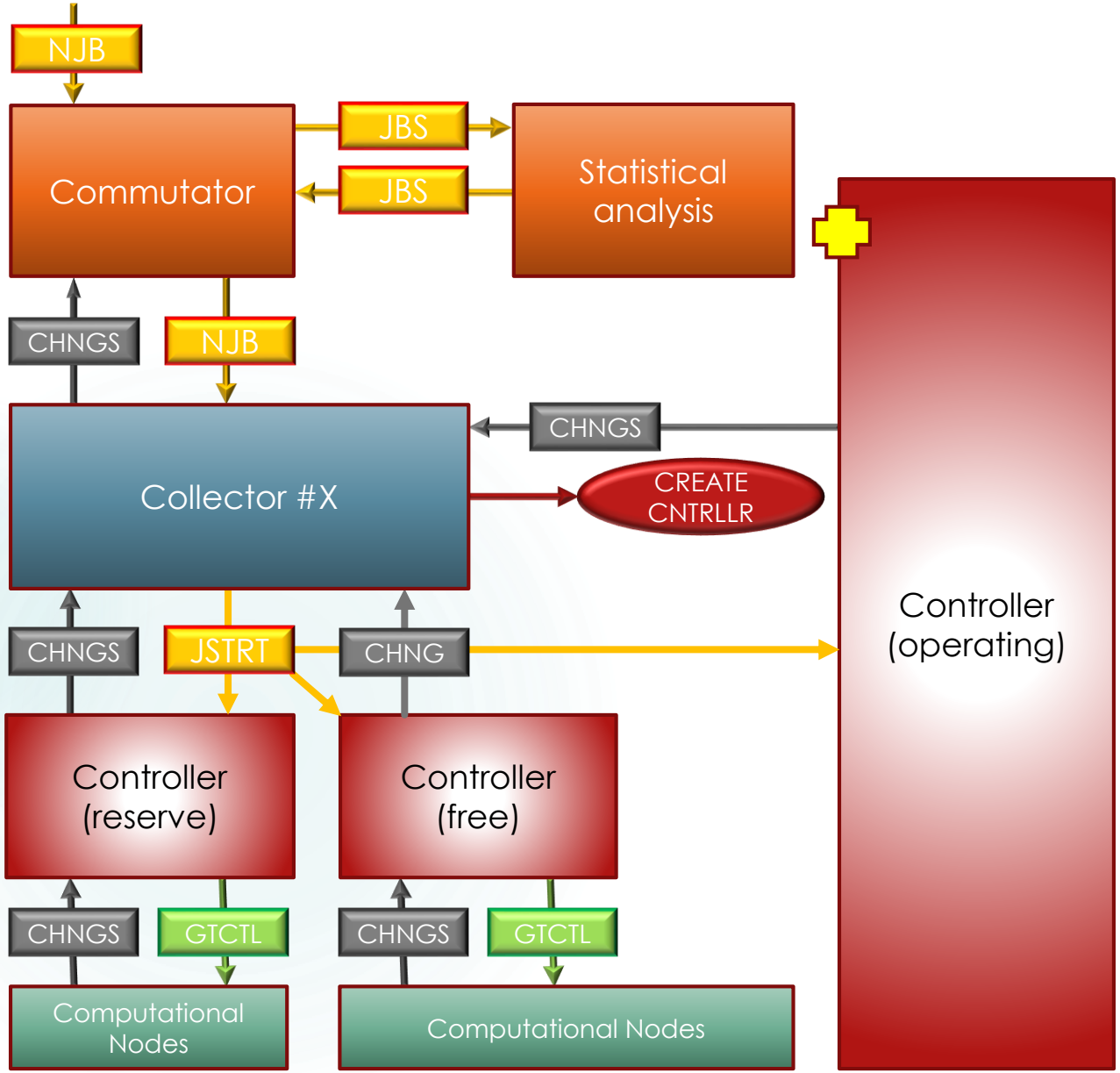
Model of HTC System. Task processing



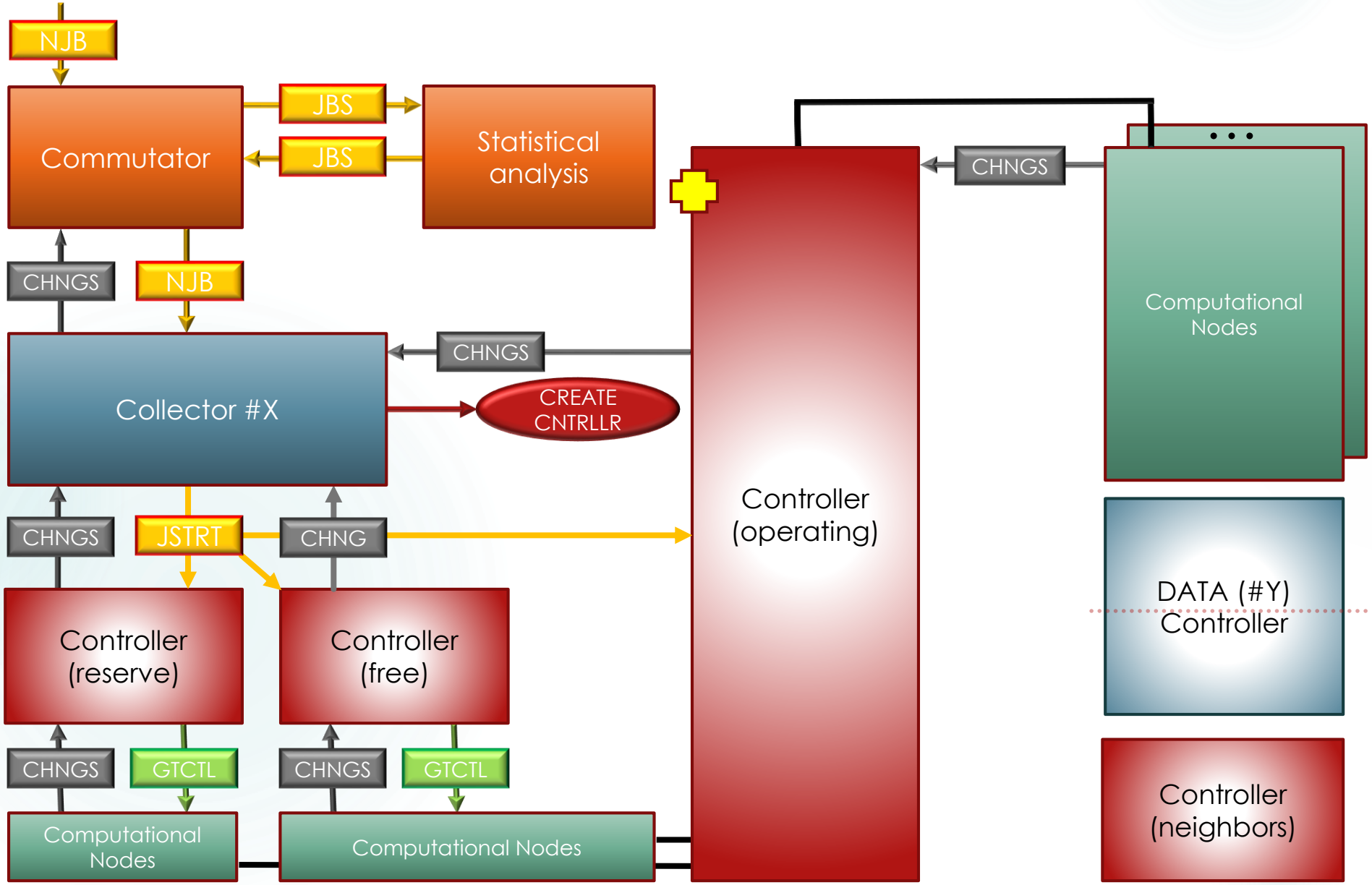
Model of HTC System. Task processing



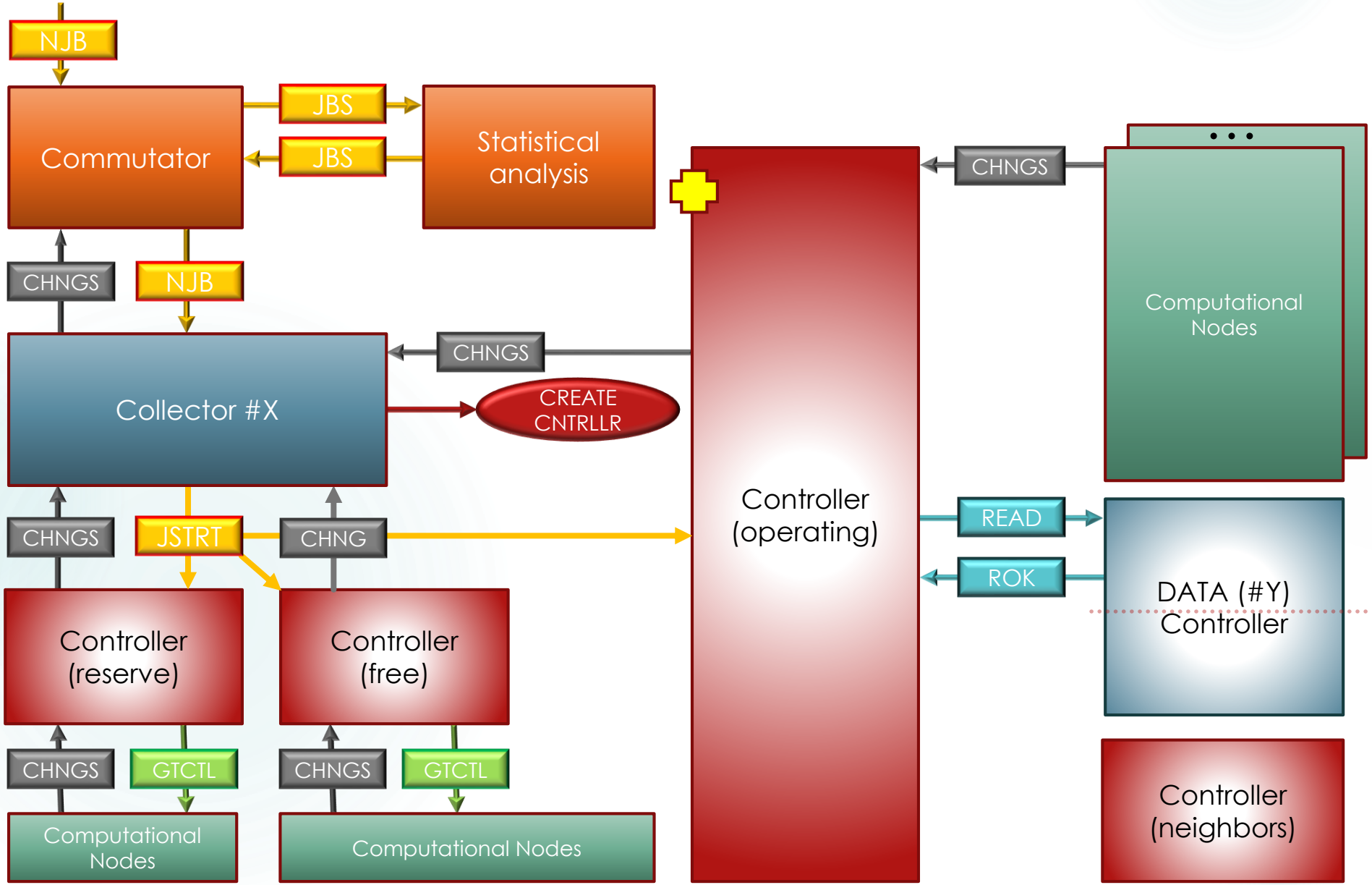
Model of HTC System. Task processing



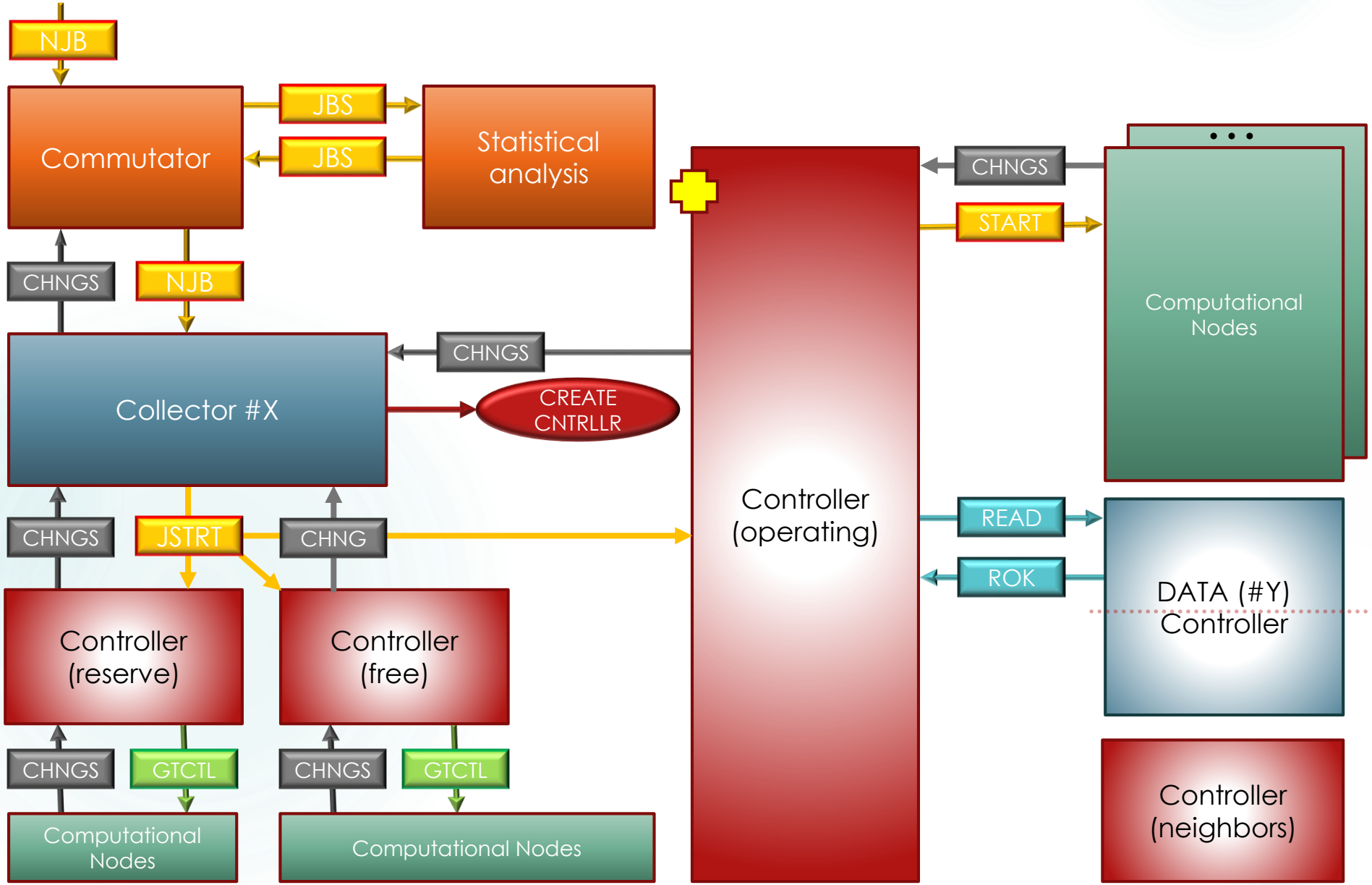
Model of HTC System. Task processing



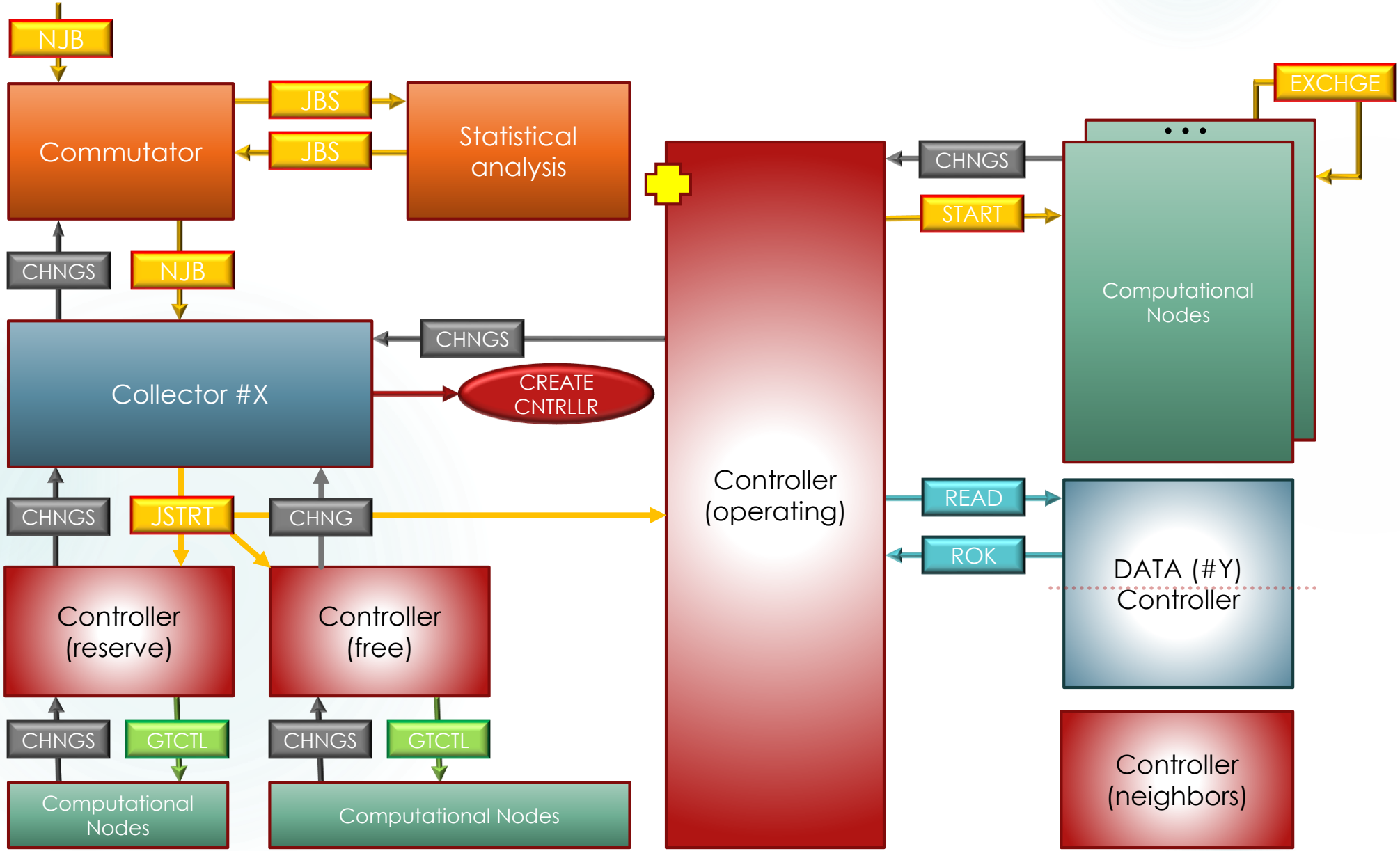
Model of HTC System. Task processing



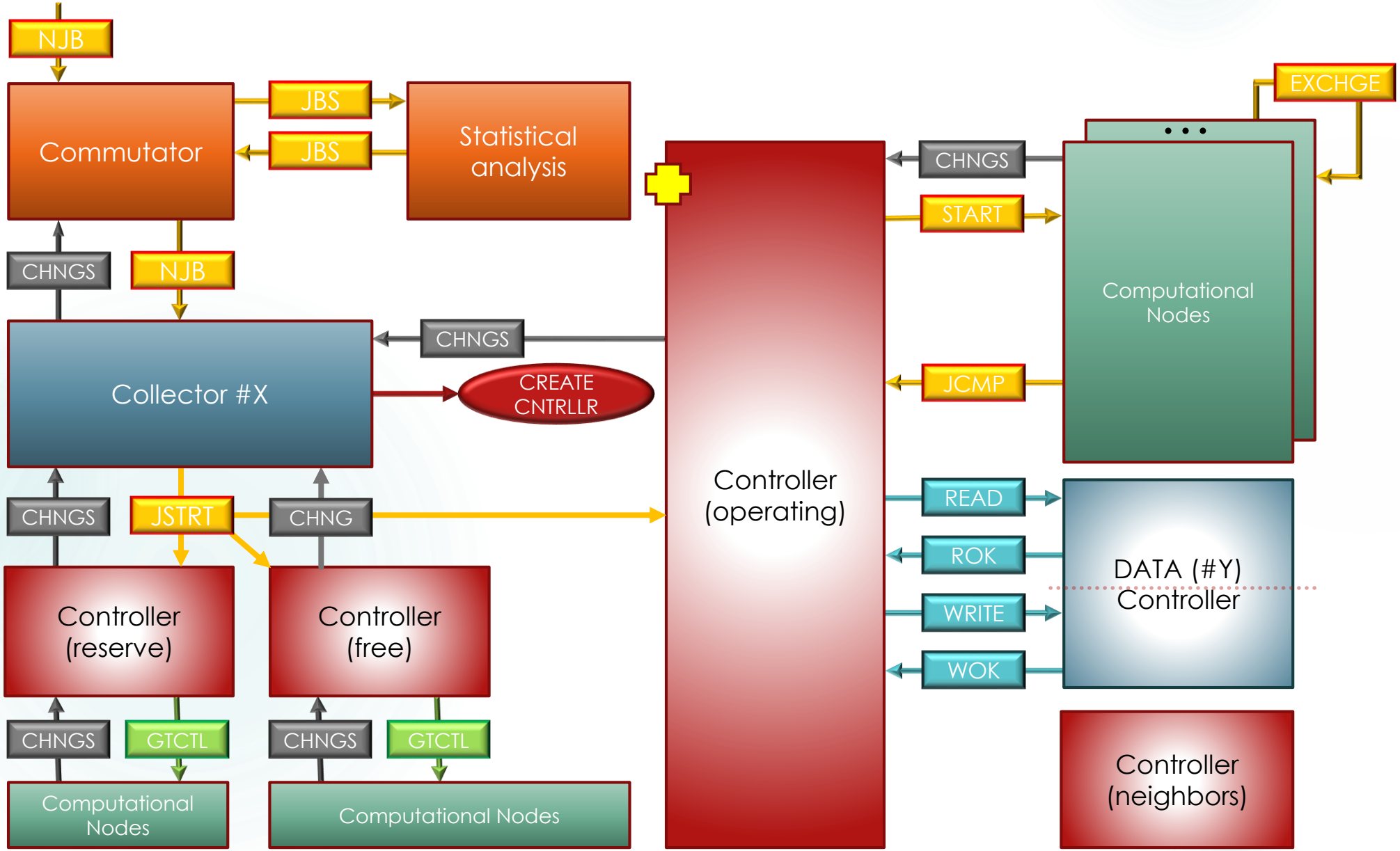
Model of HTC System. Task processing



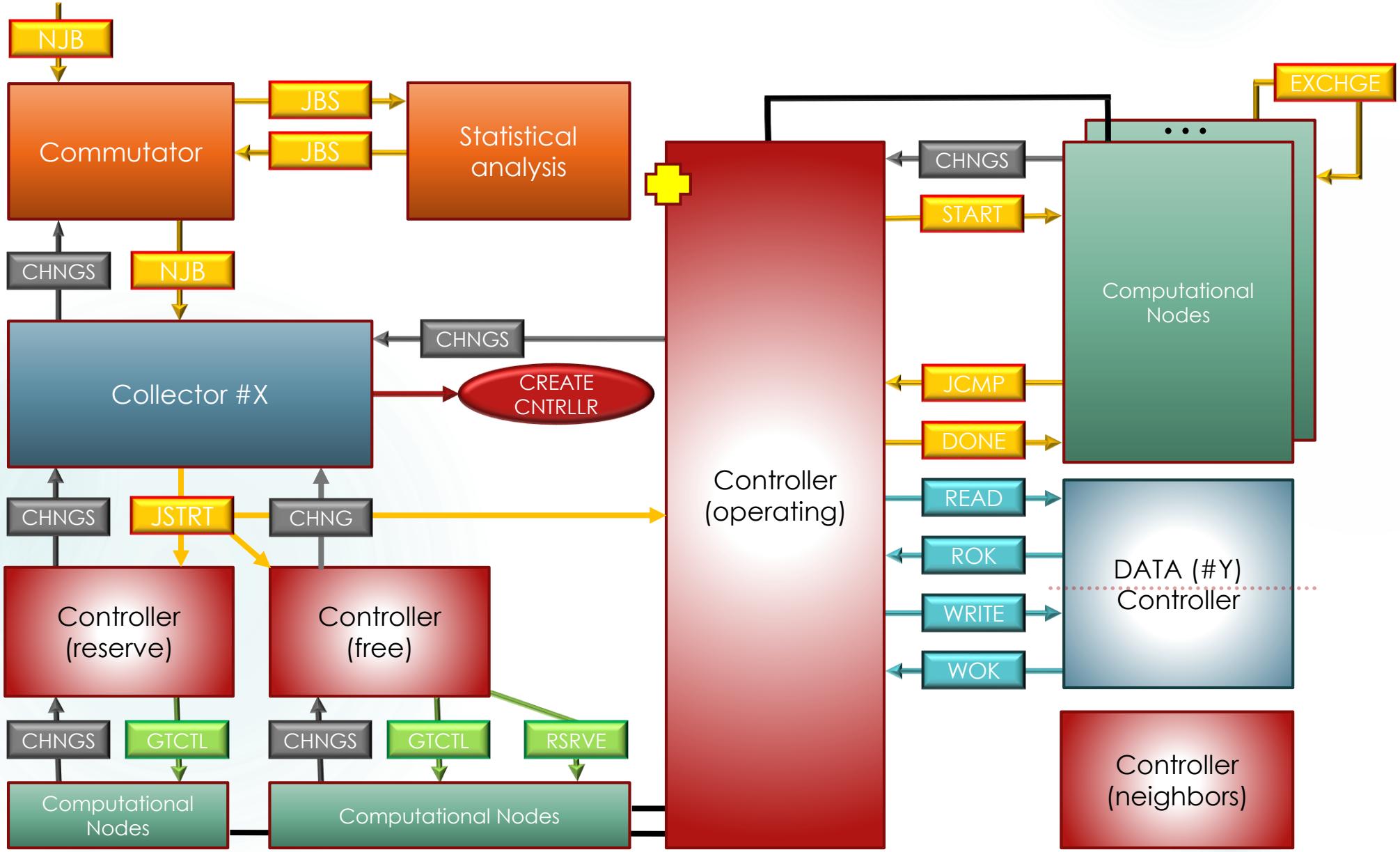
Model of HTC System. Task processing



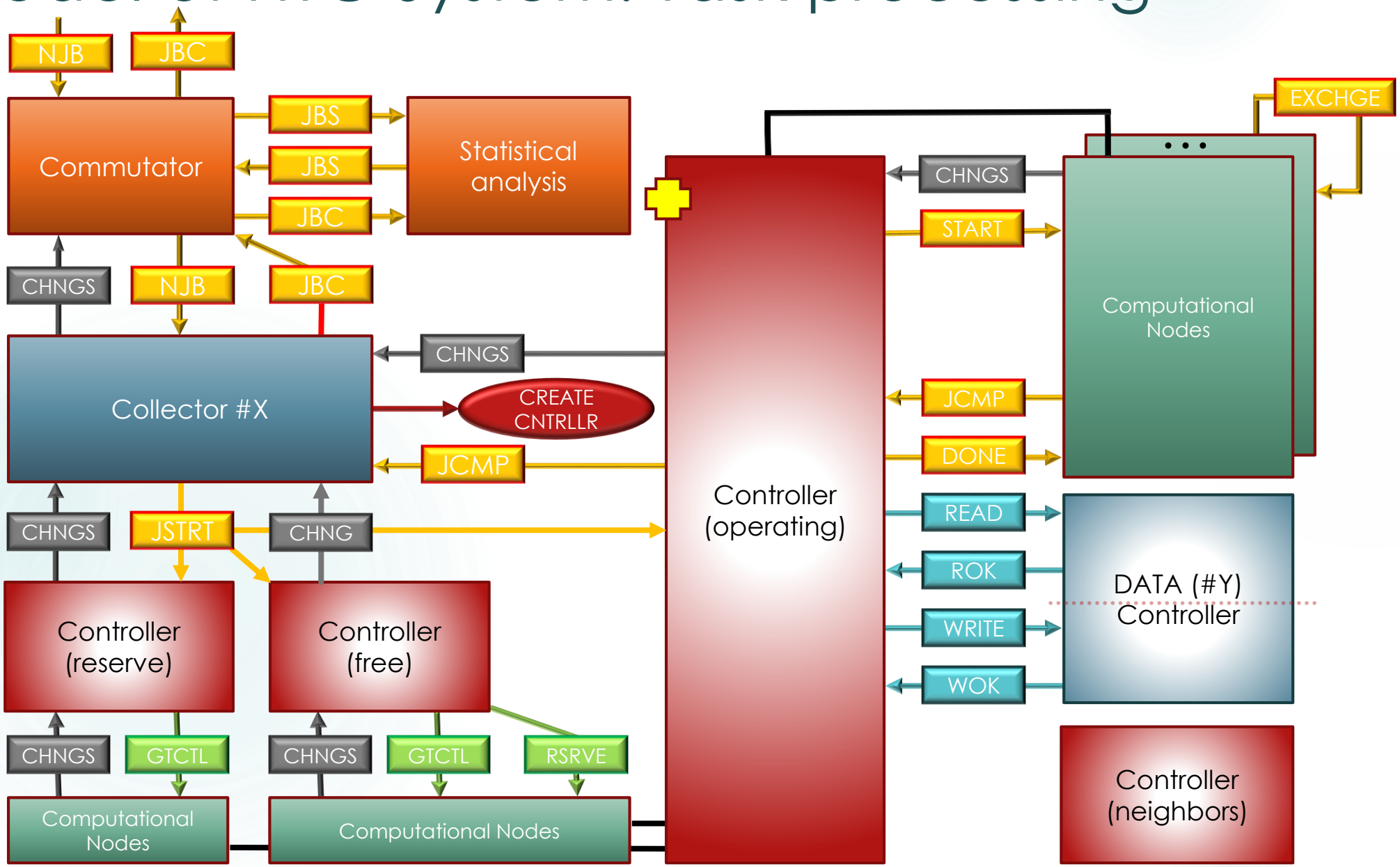
Model of HTC System. Task processing



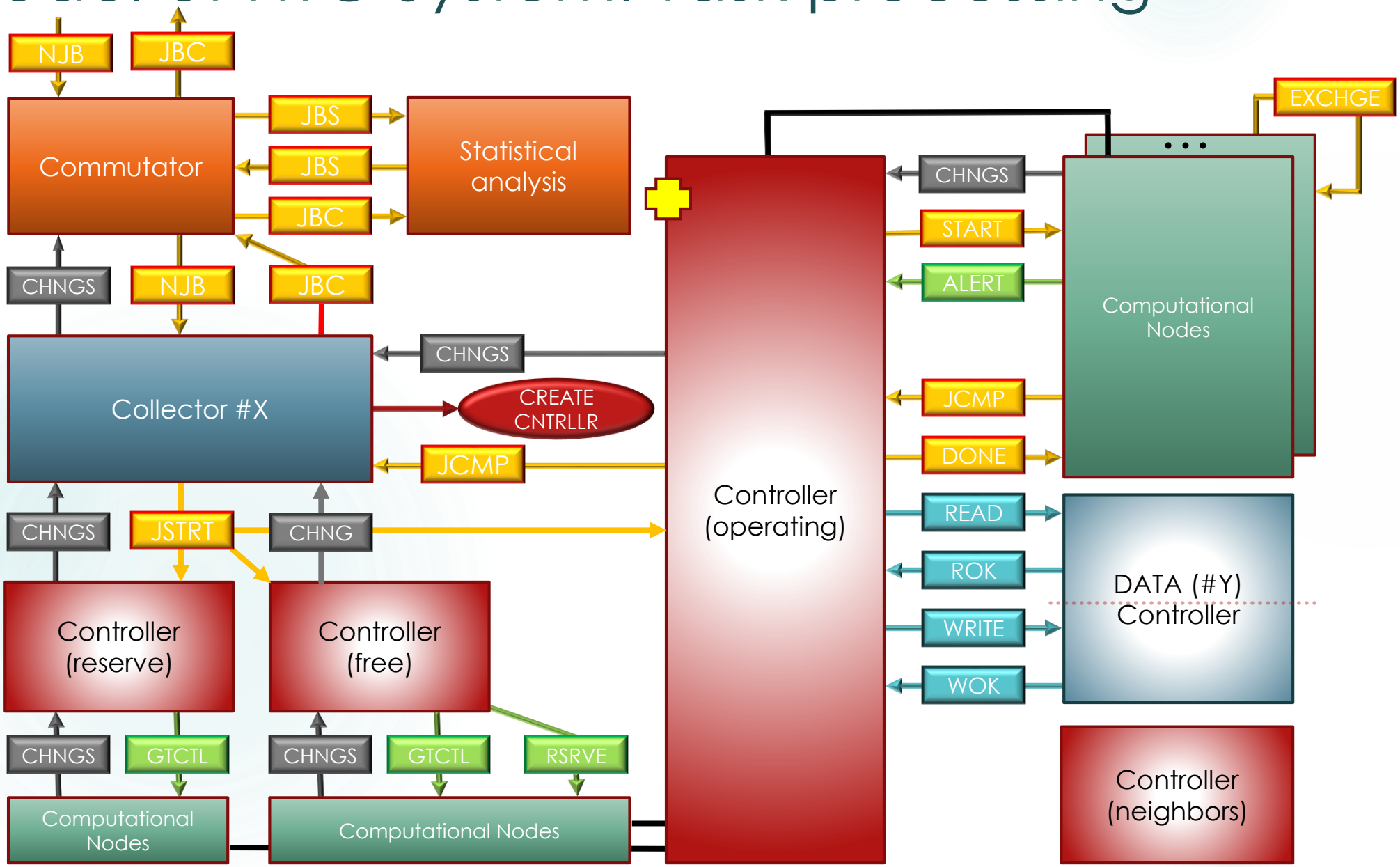
Model of HTC System. Task processing



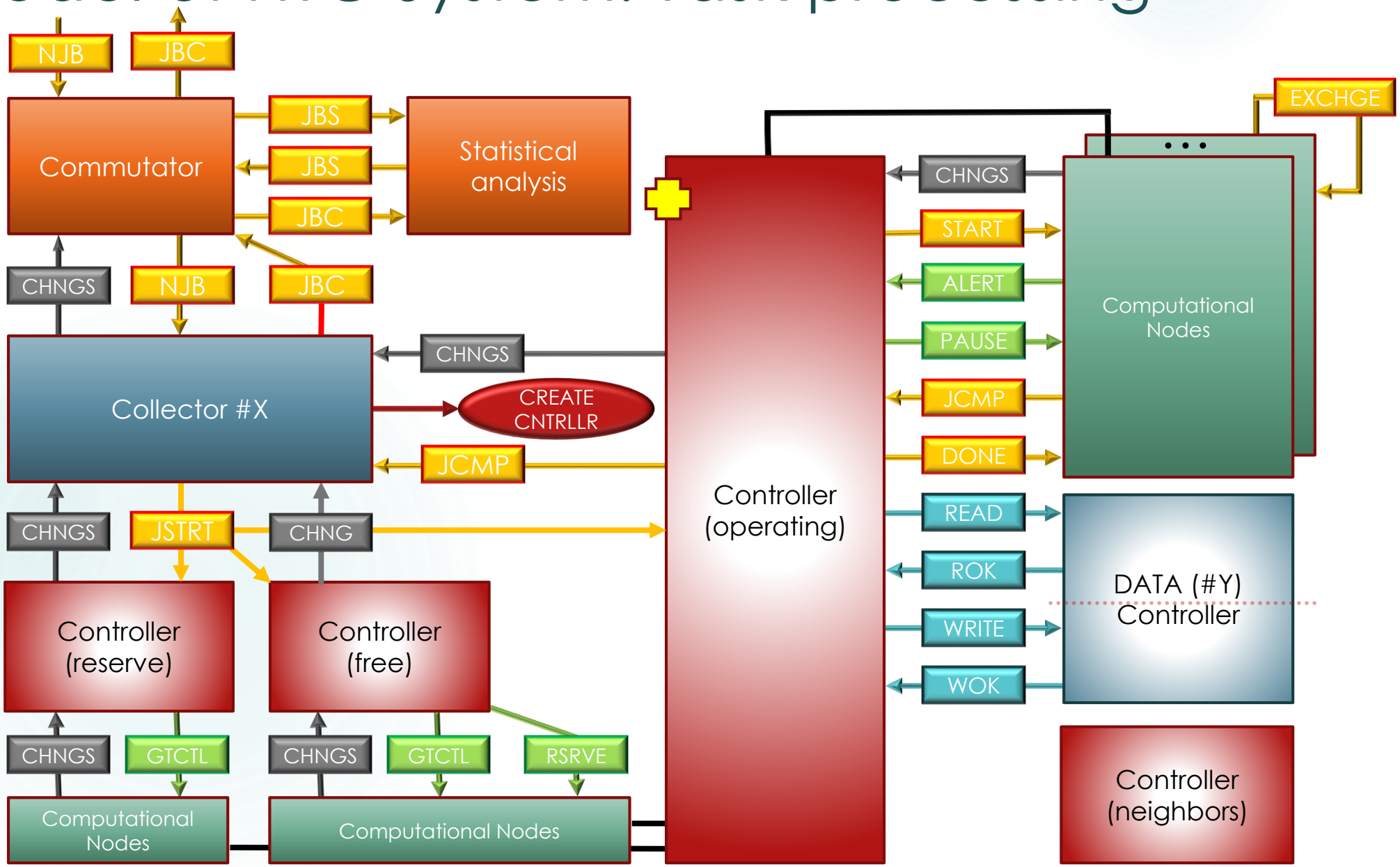
Model of HTC System. Task processing



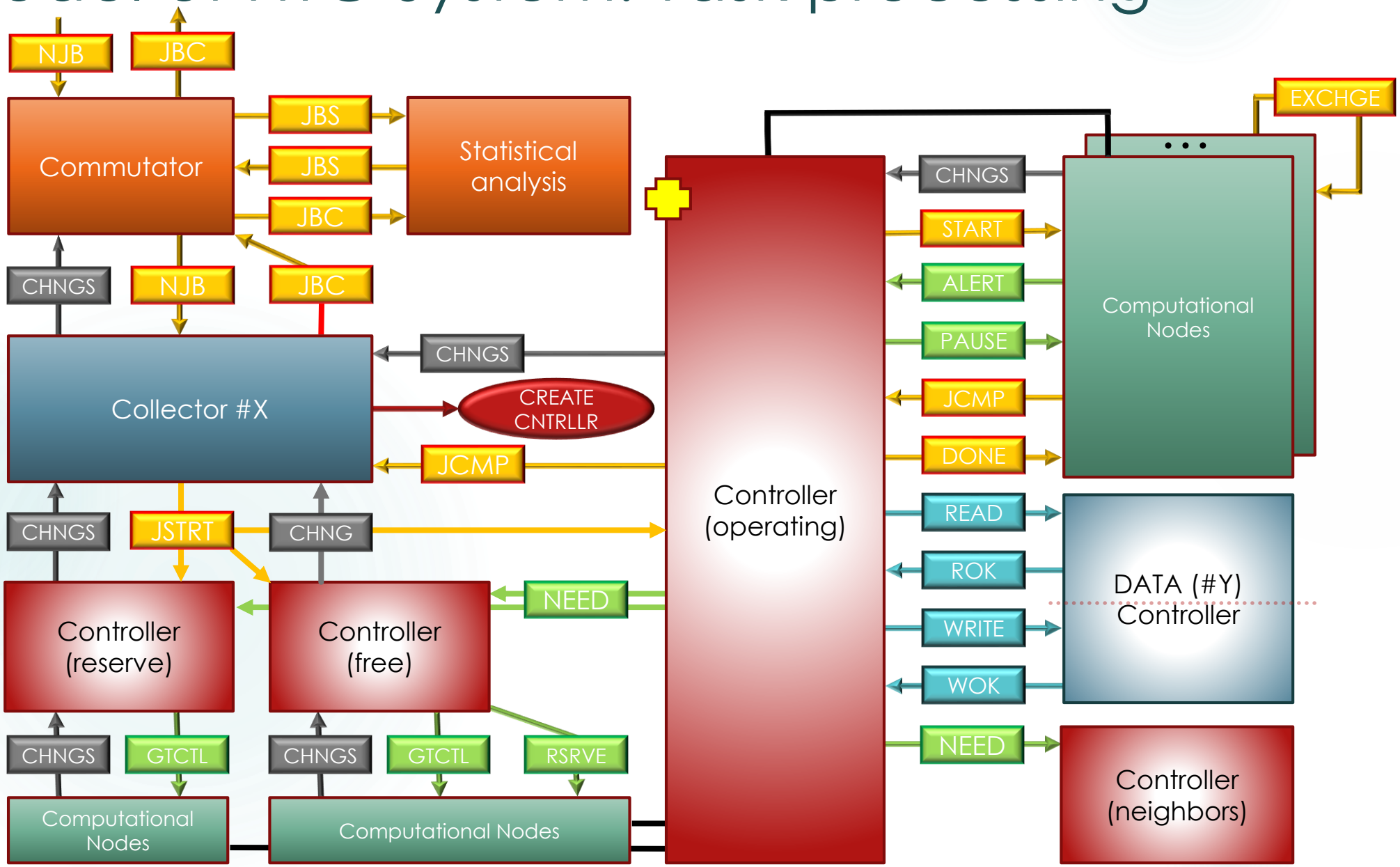
Model of HTC System. Task processing



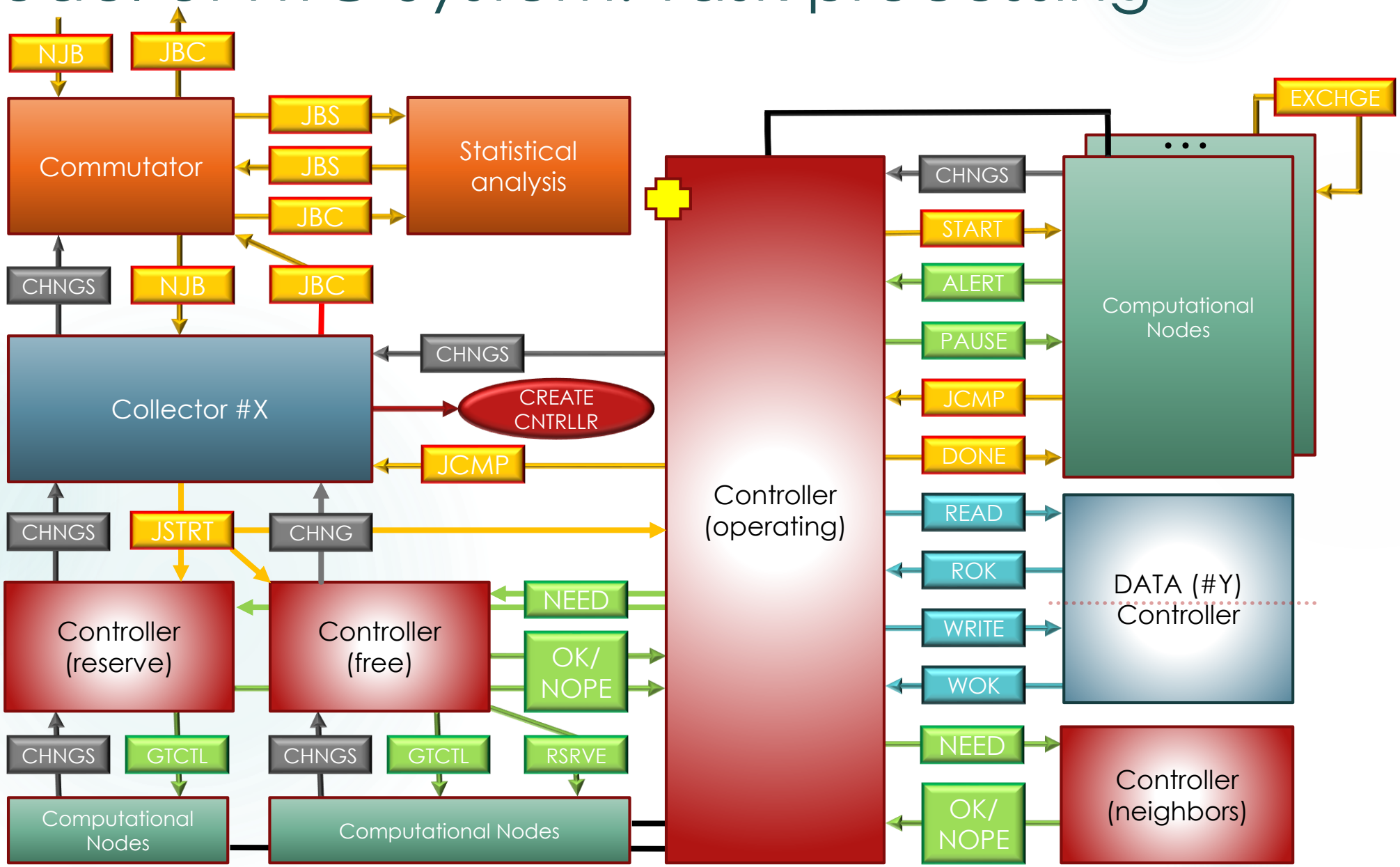
Model of HTC System. Task processing



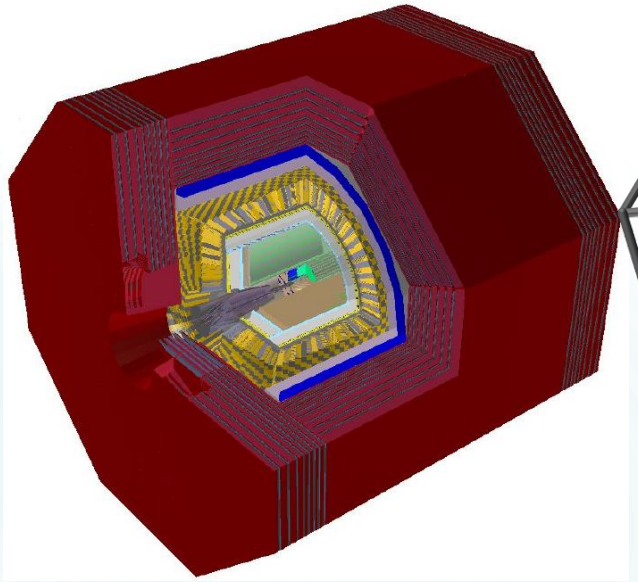
Model of HTC System. Task processing

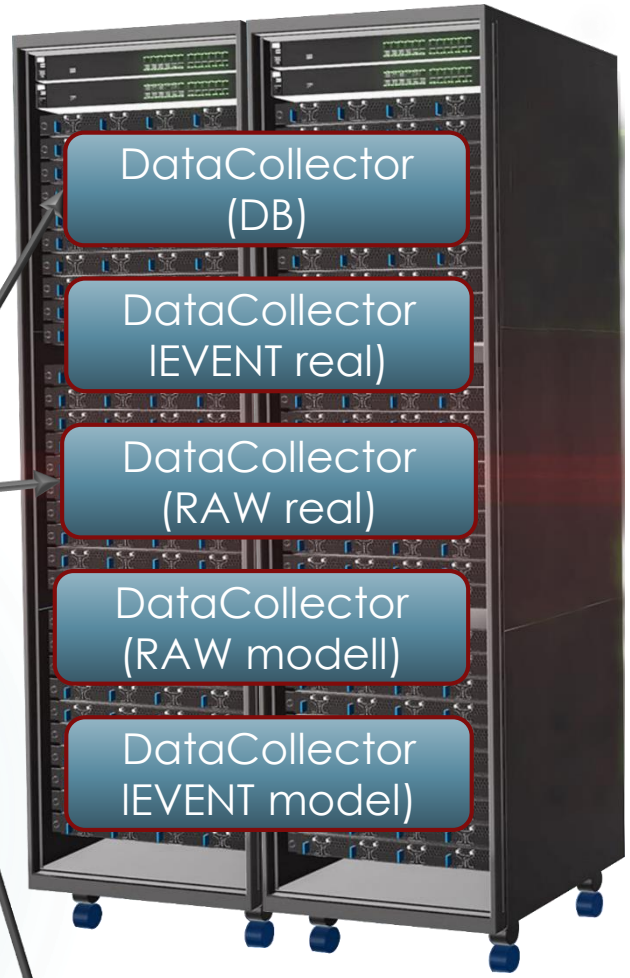


Model of HTC System. Task processing



Model of Storage Data System





DataCollector (DB)

DataCollector IEVENT real)

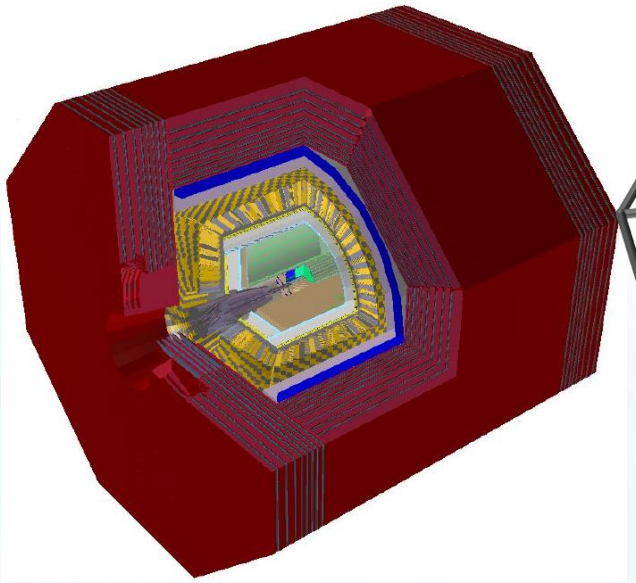
DataCollector (RAW real)

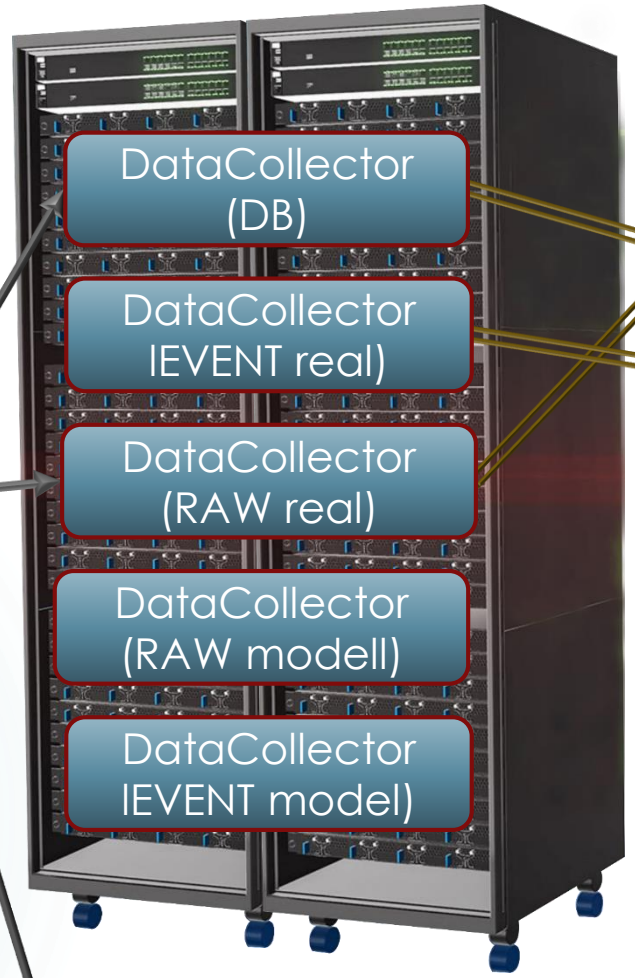
DataCollector (RAW modell)

DataCollector IEVENT model)

DataCollector (Backup)

Model of Storage Data System





DataCollector (DB)

DataCollector IEVENT real)

DataCollector (RAW real)

DataCollector (RAW modell)

DataCollector IEVENT model)

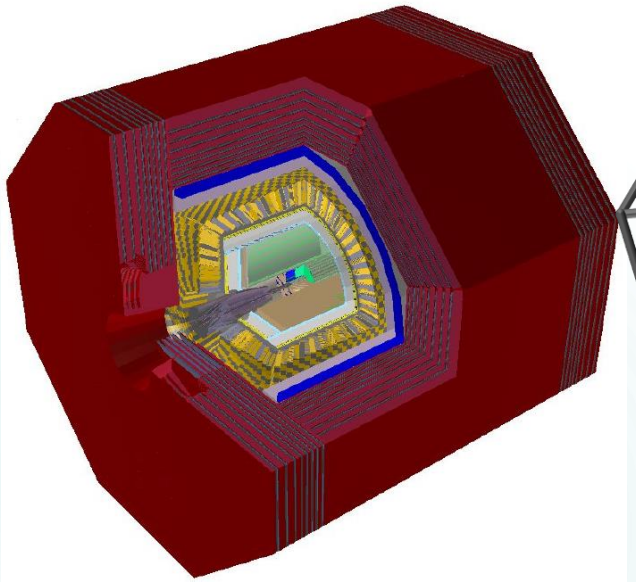
DataCollector (Backup)

Read ops

Write ops

Reconstruction task

Model of Storage Data System



DataCollector (DB)

DataCollector IEVENT real)

DataCollector (RAW real)

DataCollector (RAW modell)

DataCollector IEVENT model)

DataCollector (Backup)

Read ops

Write ops

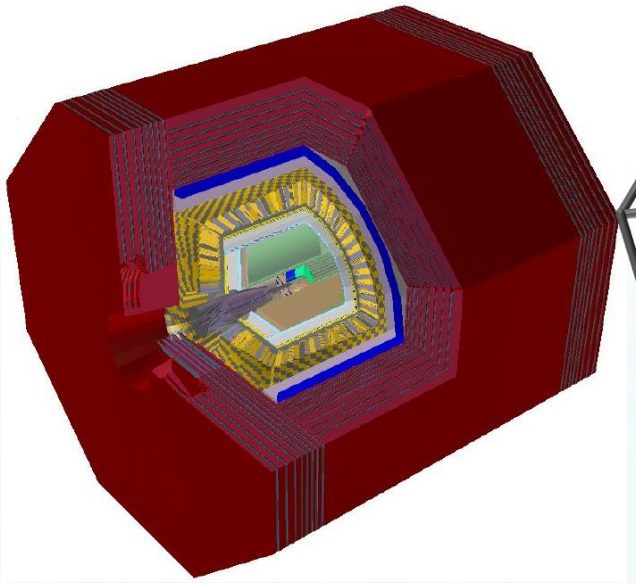
Write ops

Read ops

Reconstruction task

Modeling task

Model of Storage Data System



DataCollector (DB)

DataCollector IEVENT real)

DataCollector (RAW real)

DataCollector (RAW modell)

DataCollector IEVENT model)

DataCollector (Backup)

Read ops

Write ops

Read ops

Read ops

Write ops

Read ops

Reconstruction task

Analysis task

Modeling task

Testing modeling. NKS-1P Model..

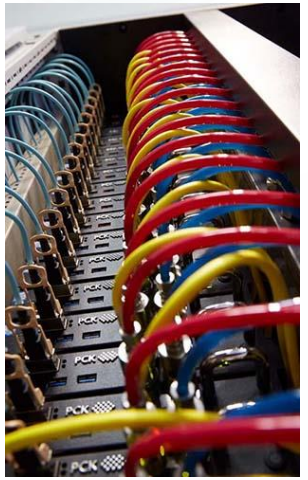
Head Node (login node)
CPU (2 x) Intel Xeon E5-2630v4 (2.2 GHz, 10 cores).
RAM 128 Gb.

(20 x) Computational nodes Broadwell
CPU (2 x) Intel Xeon E5-2697A v4 (2.6 GHz, 16 cores).
RAM 128 Gb.
(16 x) Computational nodes KNL
CPU (1 x) Intel Xeon Phi 7290 KNL
(1.5 GHz, 72 cores, 16Gb cache MCDRAM).
RAM 96 Gb
Peak performance - 81,9 ТФЛОП/С

Parallel File System
Intel Lustre - 200 ТБайт.

+ LIH SB RAN segment
(7 x) Computational nodes Broadwell
CPU (2 x) Intel Xeon E5-2697v4 (2.6 GHz, 16 cores).
RAM 256 Gb.

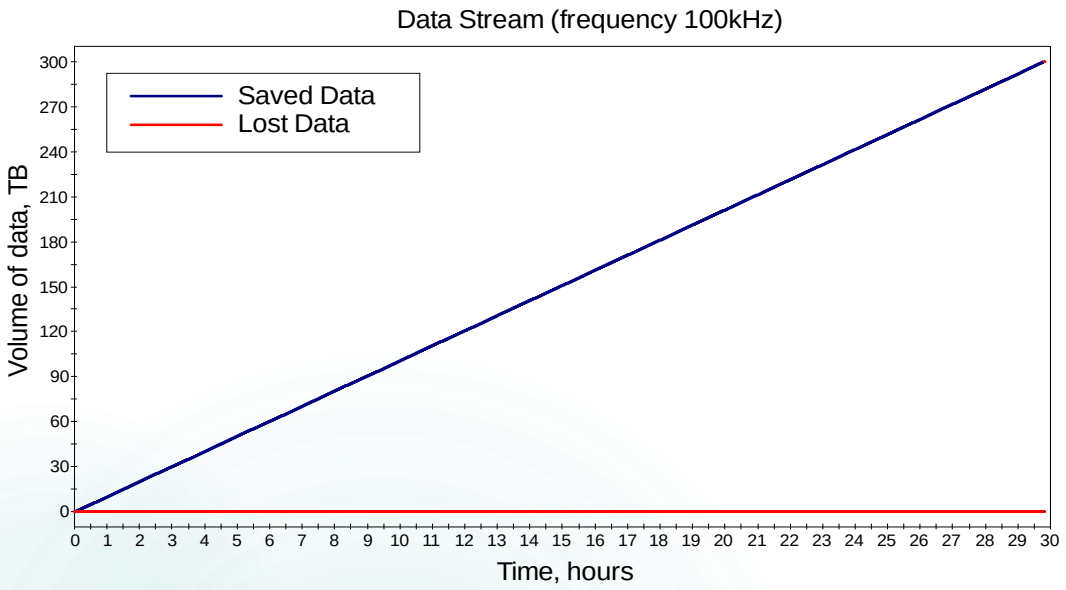
Storage Data System- 100ТБ

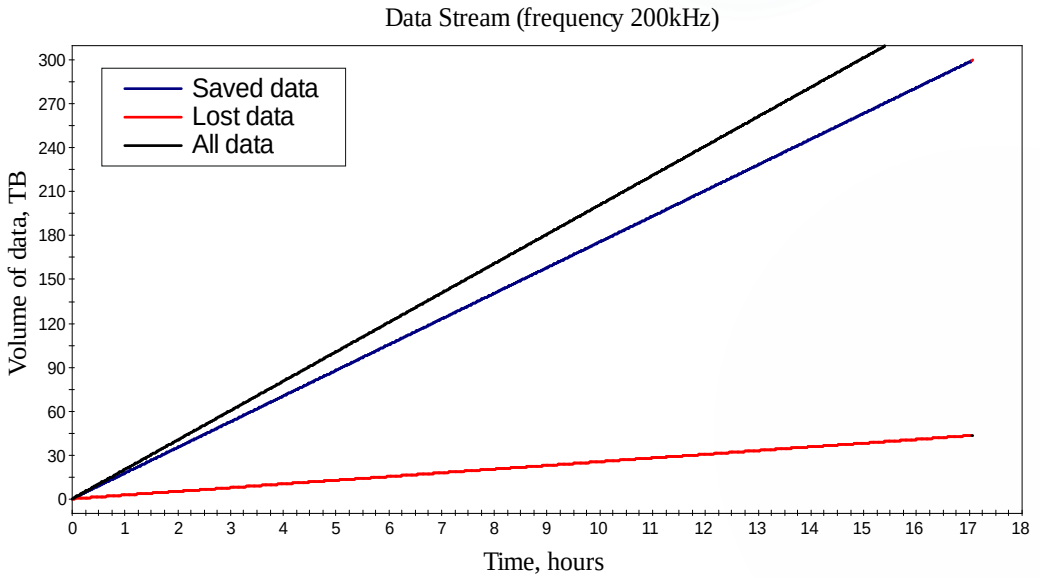


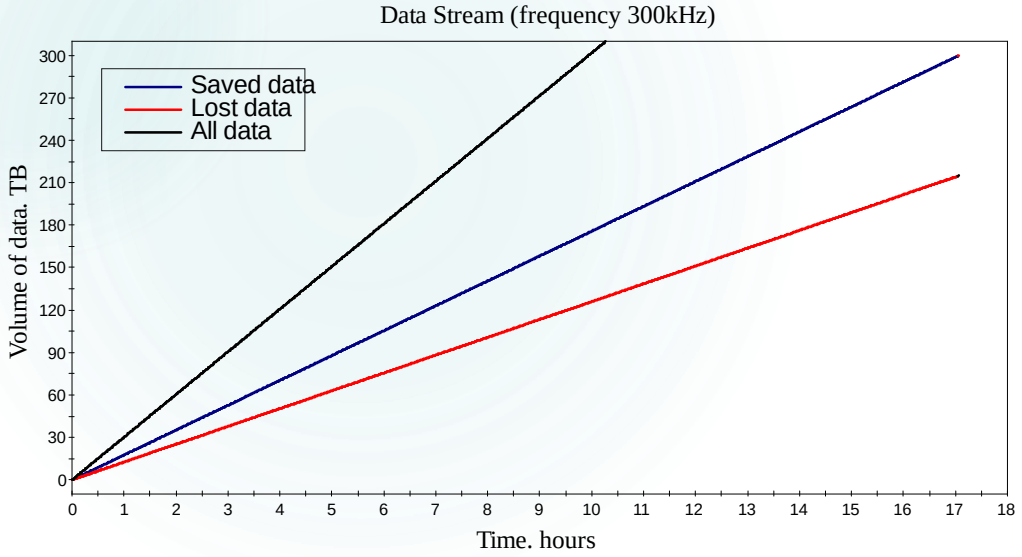
NKS1P.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<AgnesConfig nodesNum="1">
    <Agent className="User">
        <Param count="1" freq="50" p-model="20" p-analize="40" p-error="22" logging="10"/>
    </Agent>
    <Agent className="Detector">
        <Param count="1" freq="200" m-size="35" m-size-m="5" logging="0"/>
    </Agent>
    <Agent className="Detector">
        <Param count="1" freq="200" data-size="40" logging="10"/>
        <Param count="1" freq="1" data-size="10" logging="10"/>
    </Agent>
    <Agent className="Commutator">
        <Param count="1" logging="10"/>
    </Agent>
    <Agent className="Statistic">
        <Param count="1"/>
    </Agent>
    <Agent className="Collector">
        <Param count="3" qsize="15" qtype="2" logging="10"/>
    </Agent>
    <Agent className="Controller">
        <Param count="6" logging="10"/>
    </Agent>
    <Agent className="Node">
        <Param count="20" freq="2600" core="16" ram="128" accel="0" break="1"/>
        <Param count="16" freq="1500" core="72" ram="96" accel="1" break="0"/>
        <Param count="7" freq="2600" core="16" ram="256" accel="0" break="5"/>
    </Agent>
    <Agent className="DController">
        <Param count="1" wspeed="10" rspeed="10" capacity="1024" logging="10" type="1"/>
        <Param count="1" wspeed="90" rspeed="96" capacity="1024" logging="10" type="2"/>
        <Param count="1" wspeed="90" rspeed="96" capacity="1024" logging="10" type="3"/>
        <Param count="1" wspeed="90" rspeed="96" capacity="1024" logging="10" type="4"/>
        <Param count="1" wspeed="90" rspeed="96" capacity="1024" logging="10" type="5"/>
        <Param count="1" wspeed="90" rspeed="96" capacity="1024" logging="10" type="6"/>
    </Agent>
</AgnesConfig>
```

Testing modeling. Data flow.

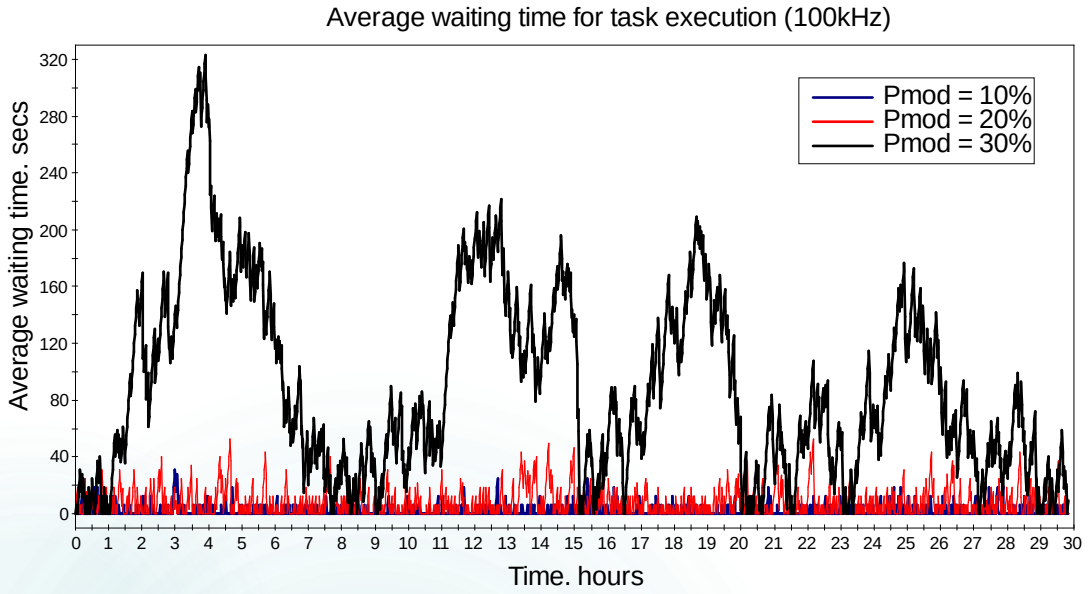


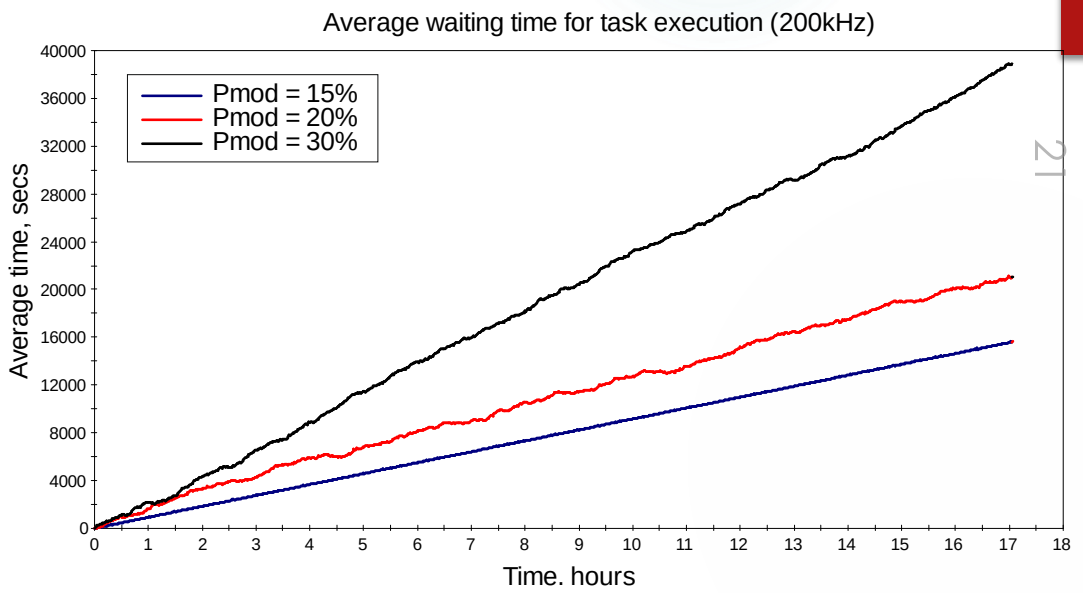


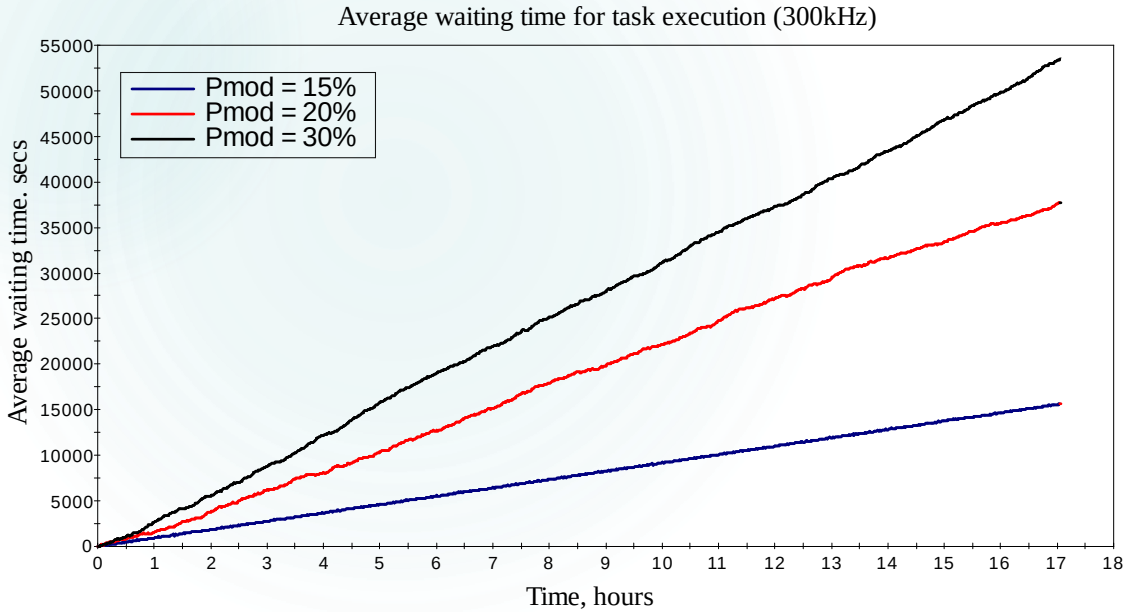


- The cluster's storage system copes with a 100kHz event stream. As the frequency increases, the volume of lost data (events) increases;
- The 300Tb data storage system will be filled within 30 hours with RAW and processed data from the detector with event frequency of 100kHz. Significantly large amounts of memory are required to store events from the detector,

Testing modeling. Task scheduling.







- The computing infrastructure copes with processing packets of events with a frequency of 100 kHz. The waiting time for task execution increases significantly with an increase in the frequency;
- Increasing the probability of receiving simulation tasks also increases the waiting time. An increase in the number of computing nodes is required.

Full-scale modeling. Description

A full-scale simulation is planned to assess the necessary and sufficient amount of computing resources to ensure the operation of the SCTF.

Full-scale modeling means the launch of a model in which the SCTF will not be limited by either disk space or computing nodes. The number of resources required at the moment of the model time will be added automatically.

Thus, at a certain point in time, the number of computing nodes will reach not only the necessary number for processing the expected task flow, but also sufficient for a given probability of failures. Of course, the required amount of memory will constantly increase, but it will be possible to accurately assess the trend of this growth.

Naturally, for modeling will be used the characteristics of the most modern server equipment at the current time.

Conclusion

The successful launch of the SCTF requires estimating the computational infrastructure parameters of the complex for storing and processing data of the physical experiment at the stage of design. Using simulation modeling allows for the maximally reliable representation of the exact characteristics and volume of the needed equipment for developing the desired HPC system. The simulation model described in this paper accounts for all the aspects of operation of this system from parallel data storage system to arrangement of the parallel launch of tasks. The developed system for processing software errors and equipment failures, as well as the system for ensuring energy efficiency make it possible to estimate the needed equipment with account for all possible emergency situations.

Thus, the developed simulation model allows calculating the parameters of the computing system required for processing and storing the operation results of the Super Charm-Tau factory after its commissioning.

Thank you for attention