Coding for Economists

Advanced Session 4

Jian Cao

19 May 2025

Module Files

Google Drive Folder

Outline

Outline

- Large Data Problems

Outline

- Large Data Problems
- Efficient Format / Operation

Outline

- Large Data Problems
- Efficient Format / Operation
- Efficient I/O

Outline

- Large Data Problems
- Efficient Format / Operation
- Efficient I/O
- Parallel / Distributed Computing

Large Data Problems

Large Data Problems

- Memory-Bound

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Large Data Problems

Solution

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation
- Chunked I/O

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation
- Chunked I/O

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation
- Chunked I/O
- Parallel computing

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation
- Chunked I/O
- Parallel computing

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation
- Chunked I/O
- Parallel computing
- HPC, cloud

Large Data Problems

- Memory-Bound
  - Program takes long time moving files in/out RAM, or crashes due to out of memory
- Compute-Bound
  - Program takes long time waiting for computing cores

Solution

- Pandas, Numpy format
- Vectorized operation
- Chunked I/O
- Parallel computing
- HPC, cloud

Efficient Format / Operation

Efficient Format / Operation

- Pandas, Numpy data format

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous
  - Faster computation

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous
  - Faster computation
- Vectorized operation

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous
  - Faster computation
- Vectorized operation
  - Vector, matrix >> element loops

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous
  - Faster computation
- Vectorized operation
  - Vector, matrix >> element loops
  - Calls C++, CUDA

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous
  - Faster computation
- Vectorized operation
  - Vector, matrix >> element loops
  - Calls C++, CUDA
  - Bulk memory access

Efficient Format / Operation

- Pandas, Numpy data format
  - Homogeneous, contiguous
  - Faster computation
- Vectorized operation
  - Vector, matrix >> element loops
  - Calls C++, CUDA
  - Bulk memory access

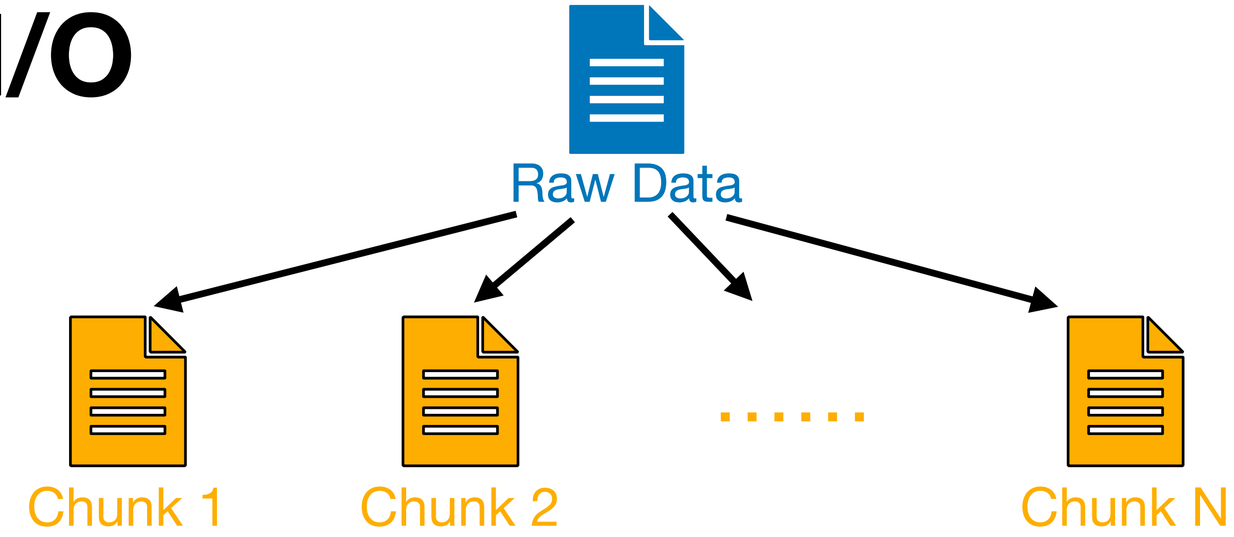
Advanced 4

Chunked I/O

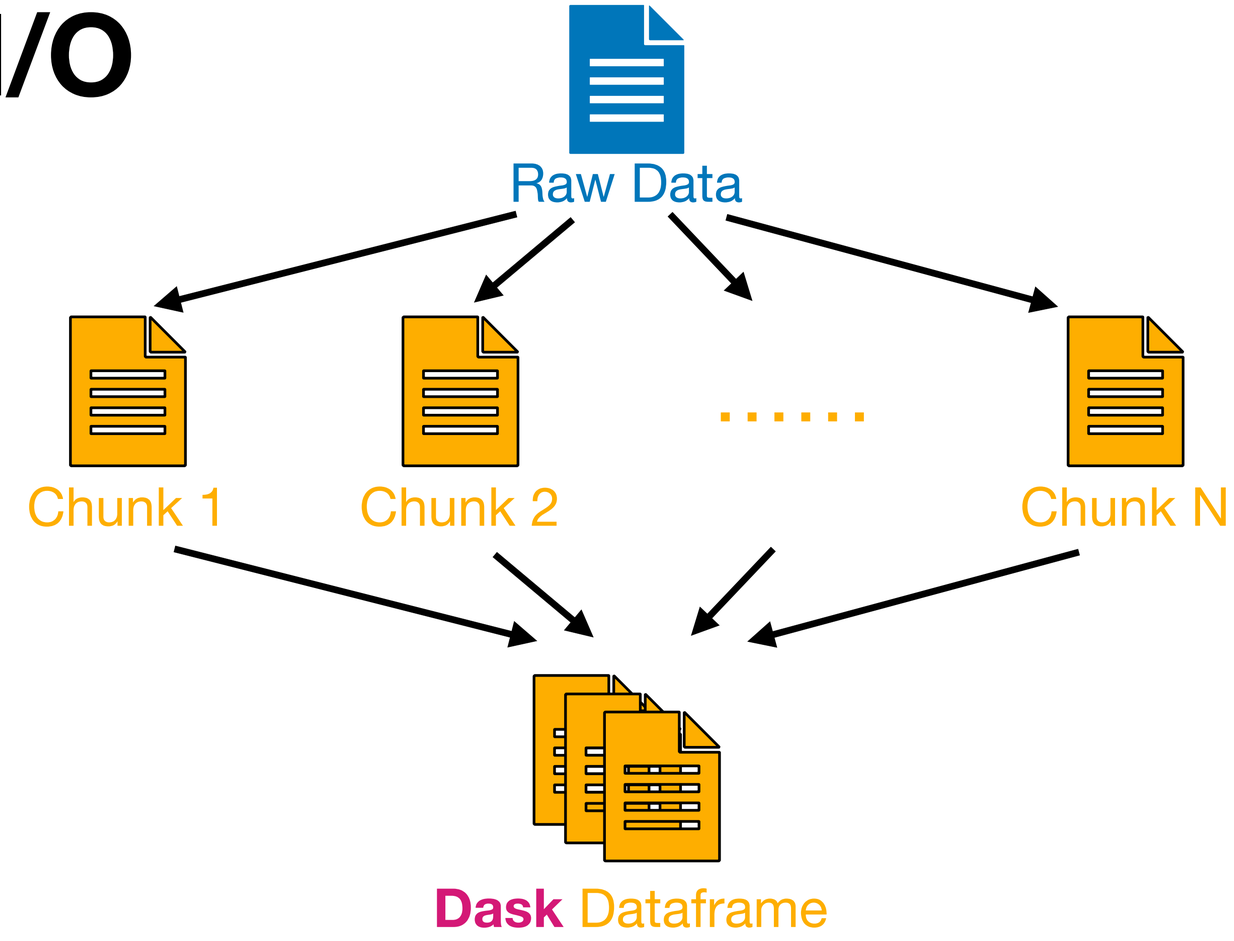
Chunked I/O

Raw Data

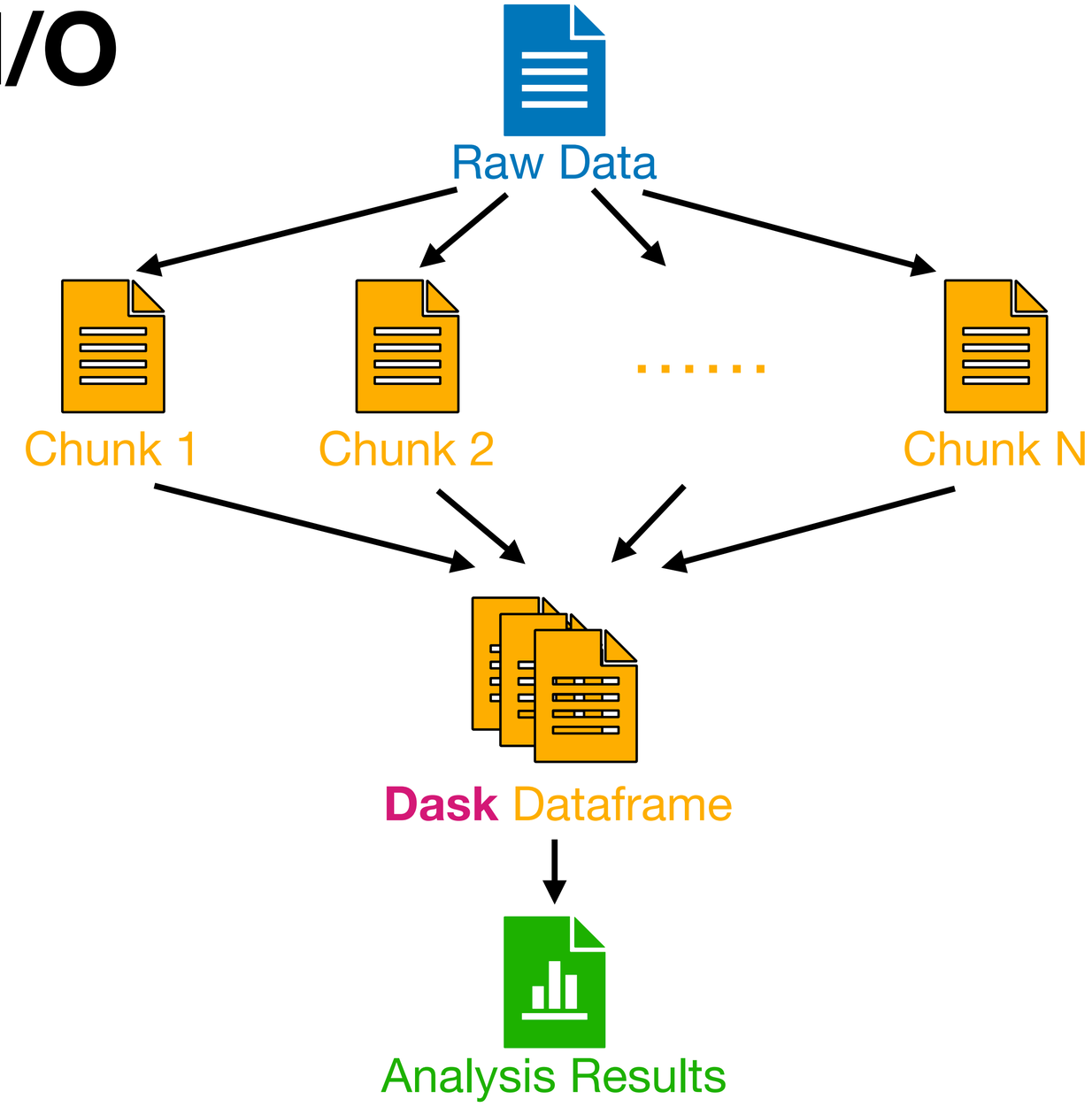
Chunked I/O



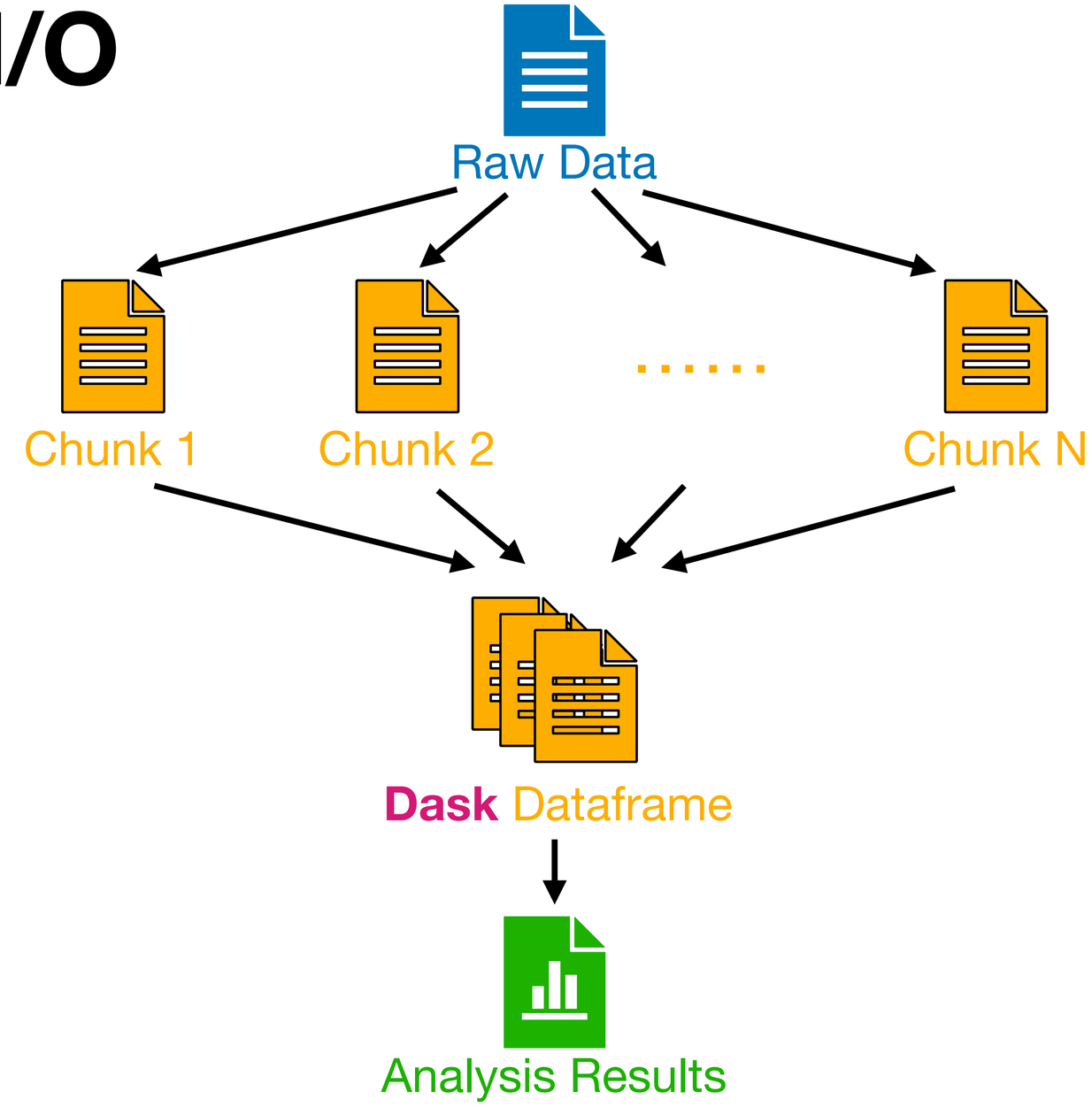
Chunked I/O



Chunked I/O


Raw Data
Chunk 1
Chunk 2
.....
Chunk N
Dask Dataframe
Analysis Results

Chunked I/O



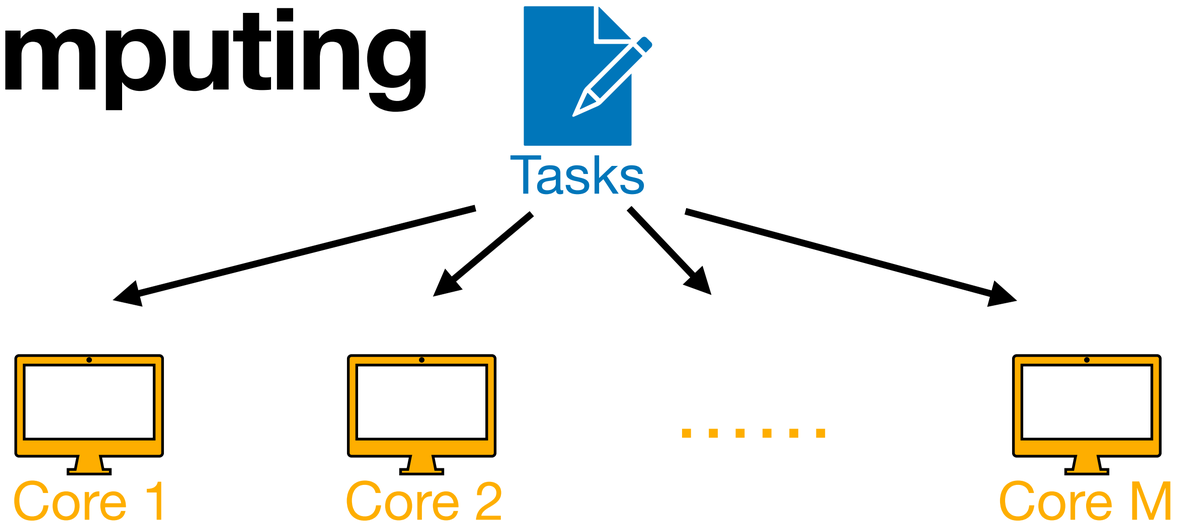
Advanced 4

Parallel Computing

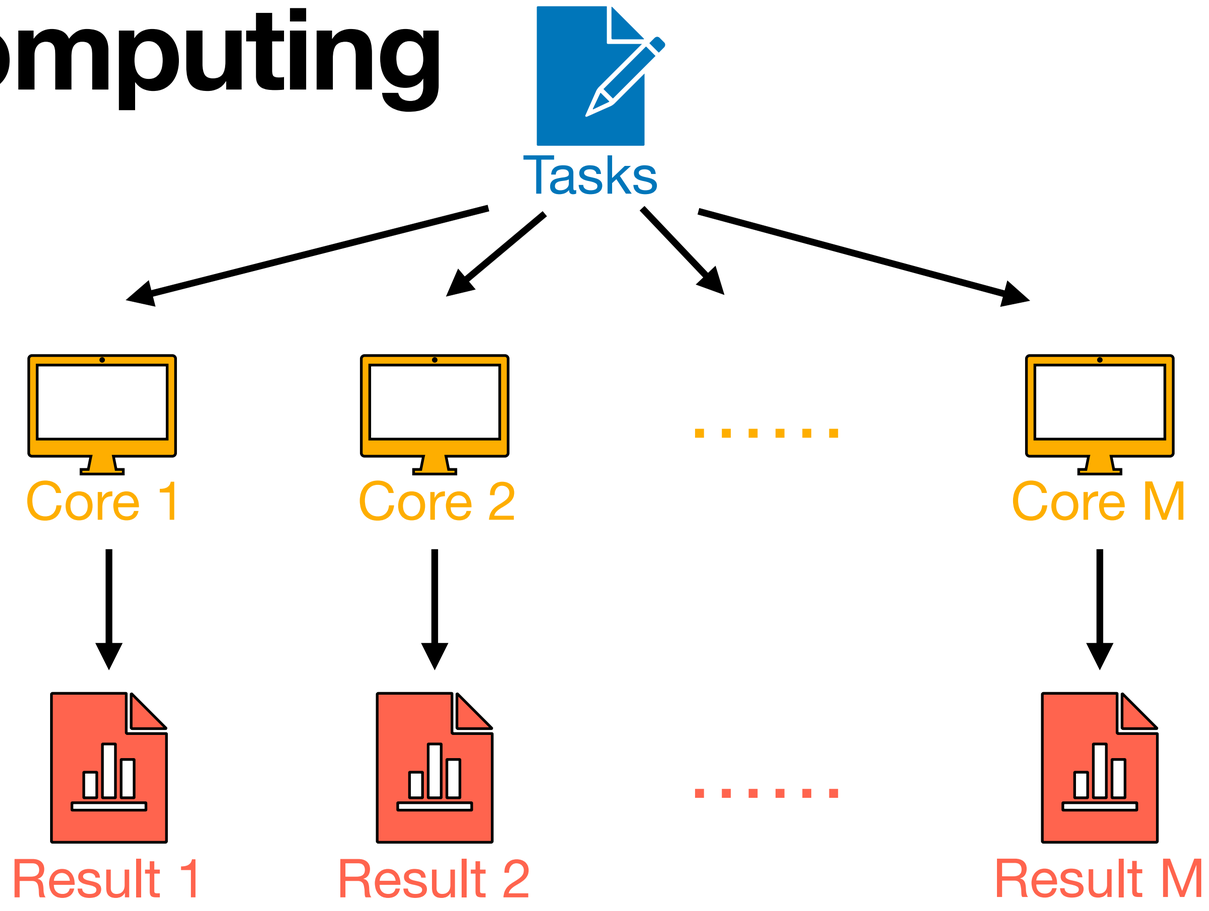
Parallel Computing

Tasks

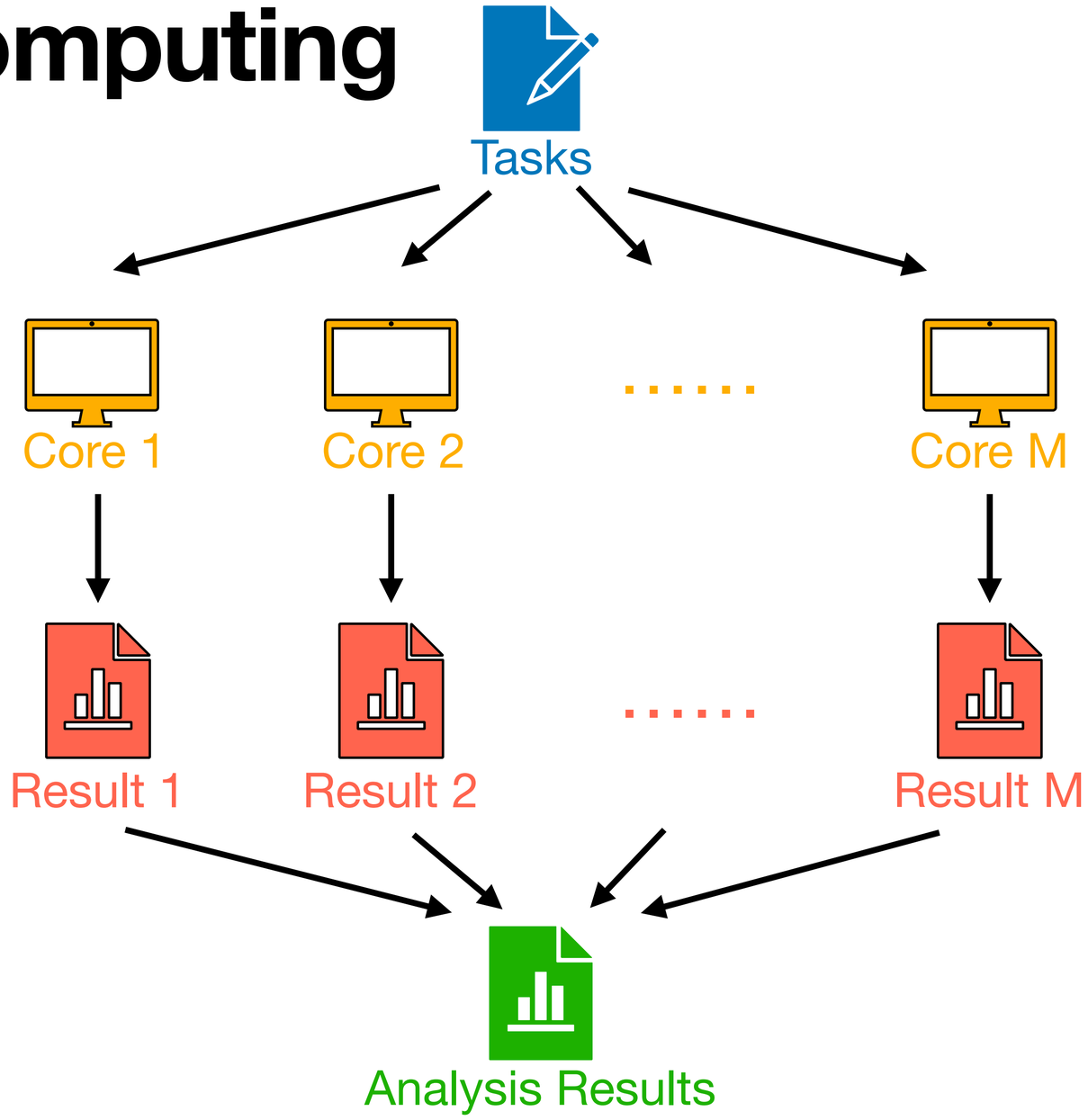
Parallel Computing



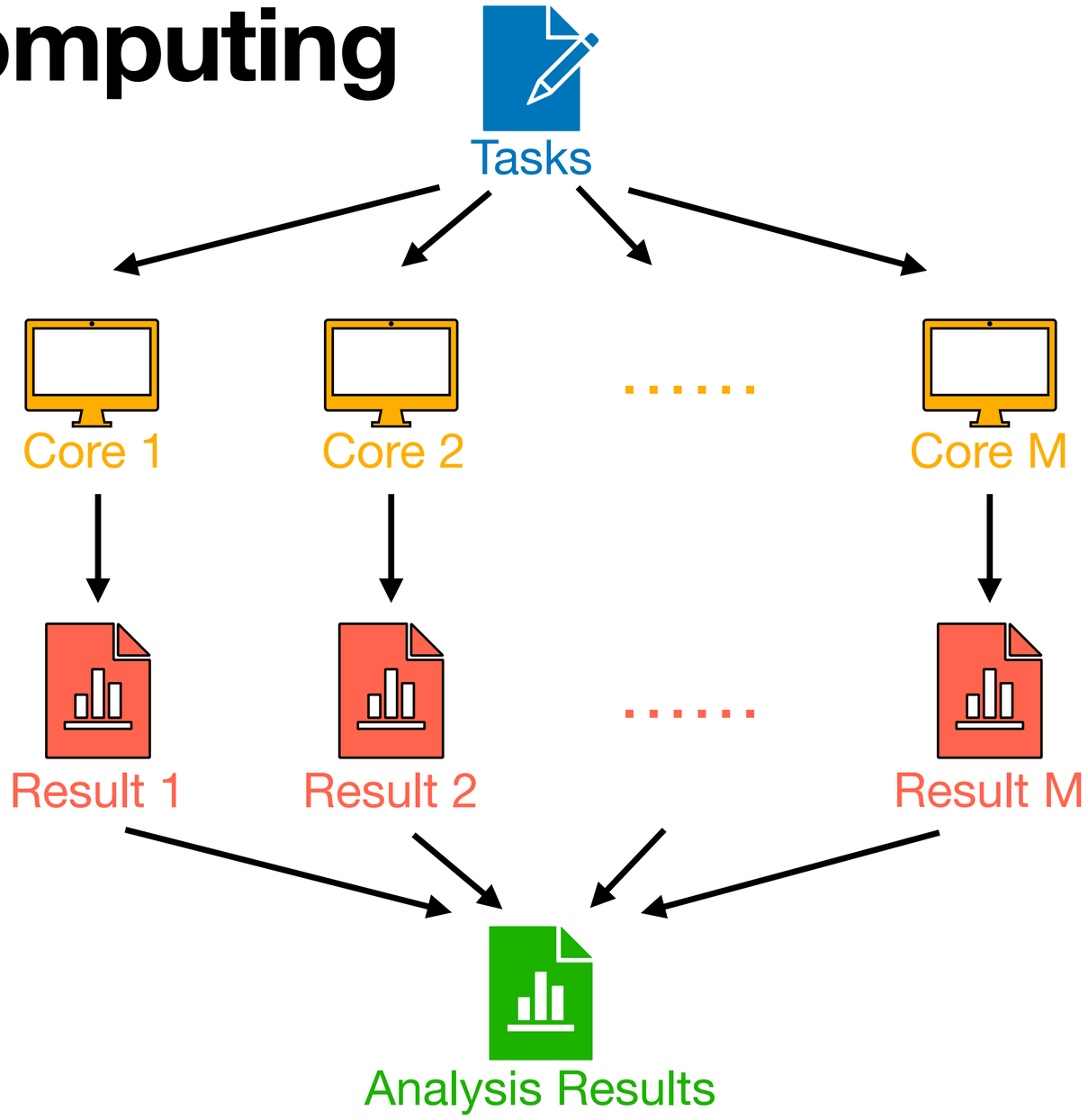
Parallel Computing



Parallel Computing



Parallel Computing



Advanced 4