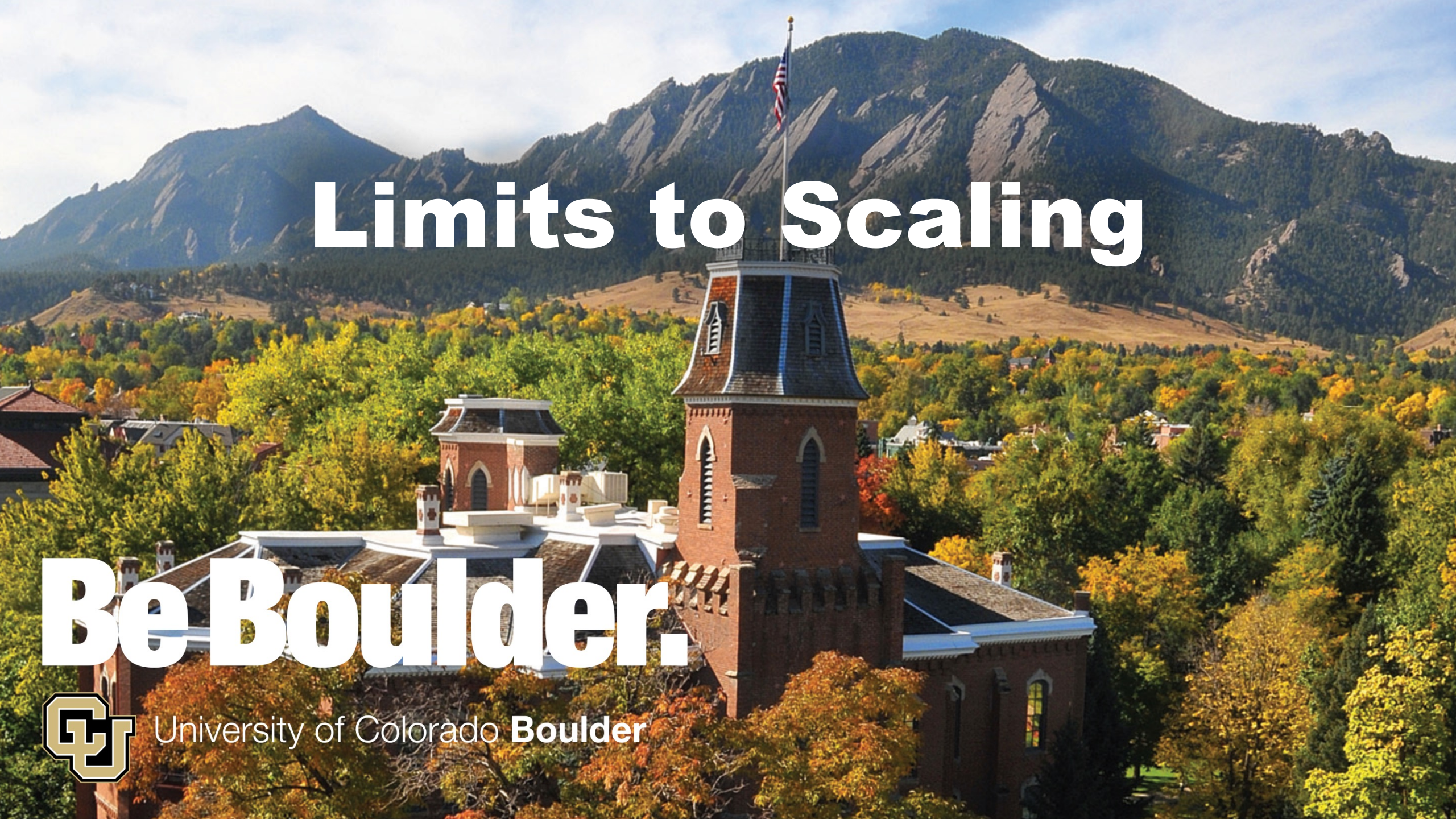

Limits to Scaling
Be Boulder.
University of Colorado Boulder

Factors that limit parallel scaling

- Ideal scaling will never happen
- Many reasons for serial portion of a parallel code
  - Algorithmic limitations, e.g. dependencies
  - Bottlenecks
  - Startup overhead
  - Communication
- Load-imbalance
- OS jitter

Algorithmic limitations

- Dependencies
  - Operations can only performed on after the other
  - Operations need to be performed in certain order

```
do i=1,nend
   a(i) = a(i-1) + a(i)    ! Dependency
enddo
```

Bottlenecks

- Generally depended on the computer system
- Shared resources as part of the computer architecture
  - Executions units in a core
  - Shared path to main memory
  - I/O devices
- Parallel access to a shared resource serializes execution
  - The application could be fully parallel, but for example one core after the other has to access a non-parallel file-system

Startup overhead

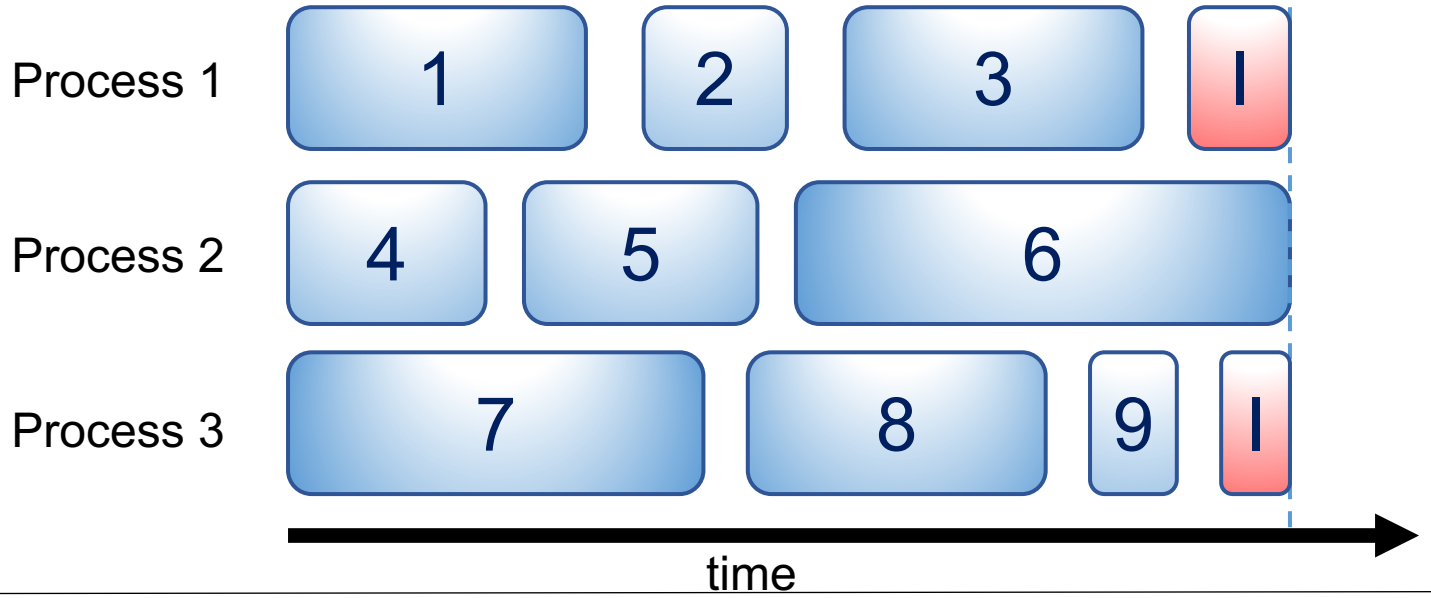
- Starting up a parallel program on a large parallel system takes time
- If run time of parallel program is too short startup time may dominate
- Similarly, shutdown overhead

Communication

- Data exchange is necessary for distributed programs
- Time spend communicating data is generally a significant source of parallel overhead
- Overhead can be reduced by overlapping communication and computation
  - Hiding the communication costs

Load imbalance

- Occurs when some workers reach synchronization points earlier than others
  - 2 processors are idling while one is still computing



OS jitter

- Particular important on large system
- Each node runs its own operating system
  - Cron jobs
  - Writing log files
  - Flushing data to disk
- OS noise will create load imbalance