

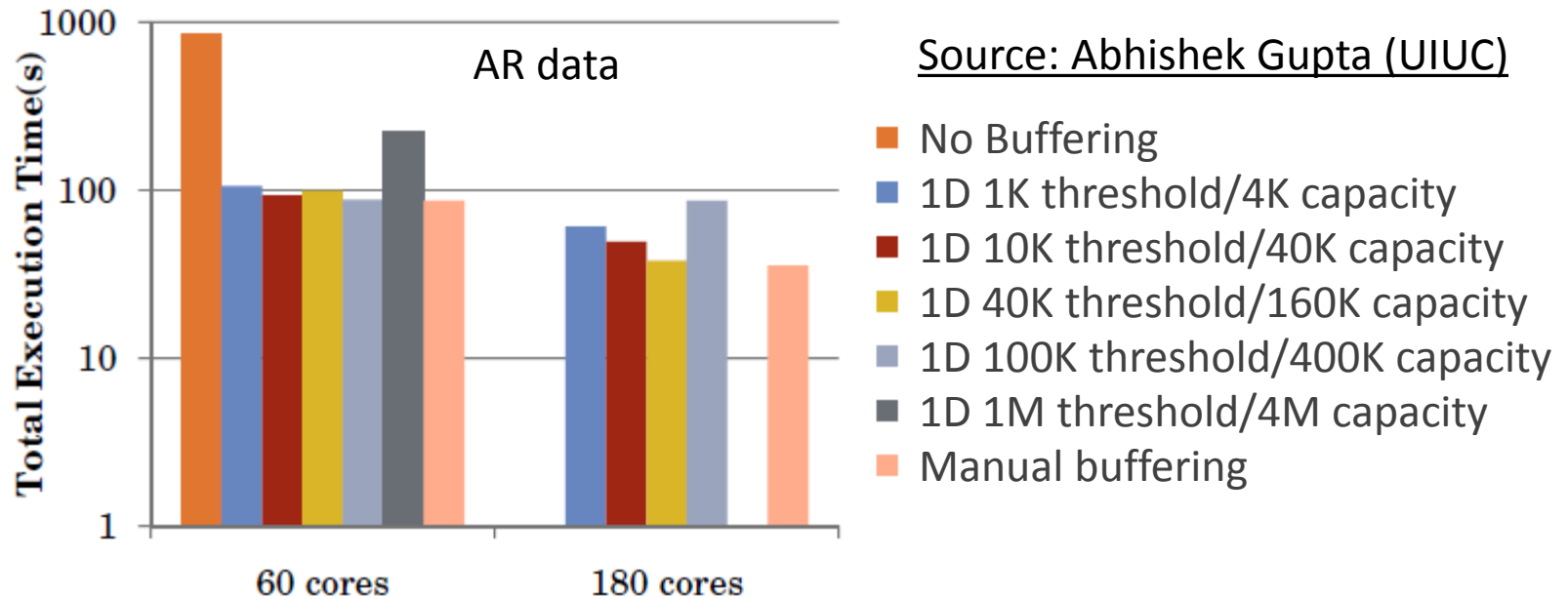
Charm++ built-in optimizations applied to EpiSimdemics

Jae-Seung Yeom

NDSSL and Computer Science

Virginia Tech

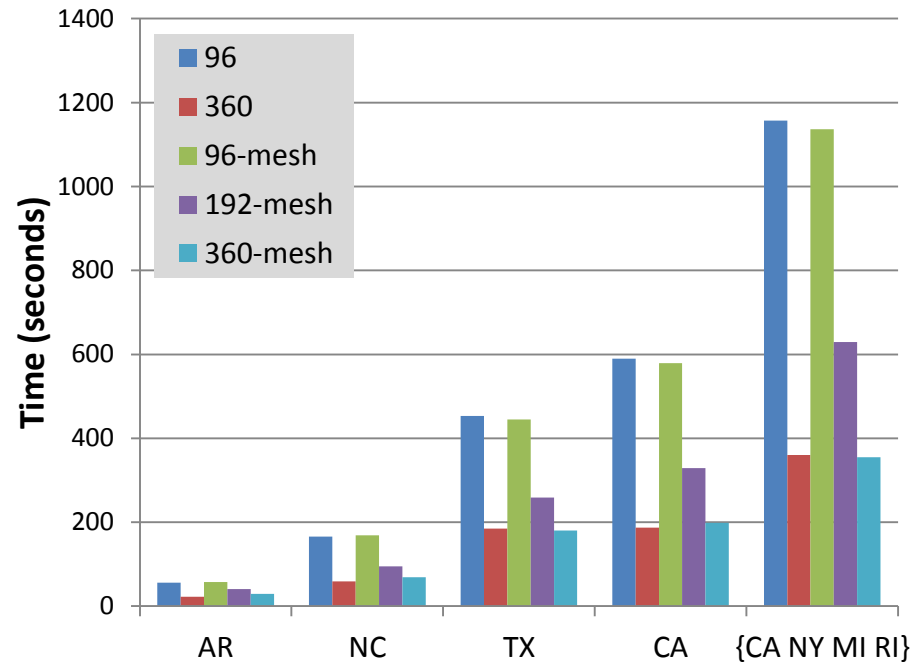
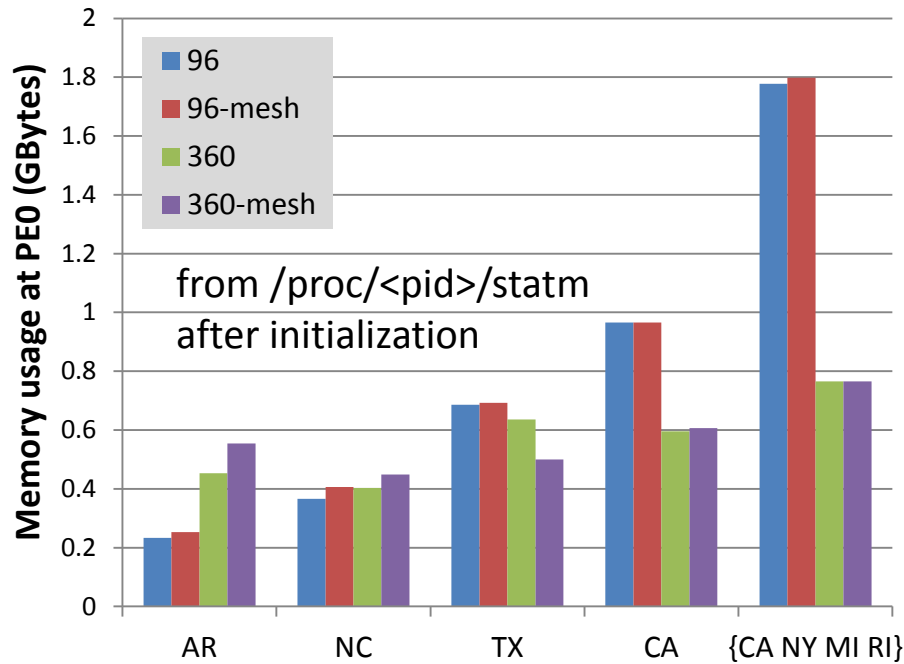
Mesh-streaming



- Mesh-streaming uses a fixed-sized total buffer and split the space (total capacity) evenly among destinations
- Mesh-streaming flushes if total number of buffered messages reaches the threshold or a buffer is full (buffer capacity = total capacity/num_buffers)
- Manual buffering uses a fixed-sized buffer per destination (Chare)

Mesh-streaming

- Relying on Group Chare
 - buffer space allocated on each PE
- Messages travel first to the correct plane, then to the column, and finally to the destination PE
- 3D topology 4x4x4 example
 - 4 buffers for PE in the row, 3 buffer to the other columns, and 3 buffers to the other planes. Total 10 buffers
 - If threshold is 40K, the total capacity is 160K (=4x40K)
 - All buffers are the same size
 - Size of each buffer = $160K / 10$
- If 10,000 message manual buffer capacity and 1 Chare per PE
 - equivalent to threshold ($2,500 * \text{NumPE}$) in terms of capacity
 - however, additional aggregate threshold is $2,500 * \text{NumPE}$
- Buffer item size
 - Manual buffer: VisitData (36 Bytes)
 - Mesh-Streaming: StreamerVisitDataMsg = 48 Bytes (VisitData + 12 Bytes)



- Mesh-streaming : 1D topo / 40K threshold / 160K capacity /48 bytes buffer unit
- Manual buffering : 10,000 messages per buffer / 36 bytes buffer unit
- 1 Chare per PE, 12 PEs per node

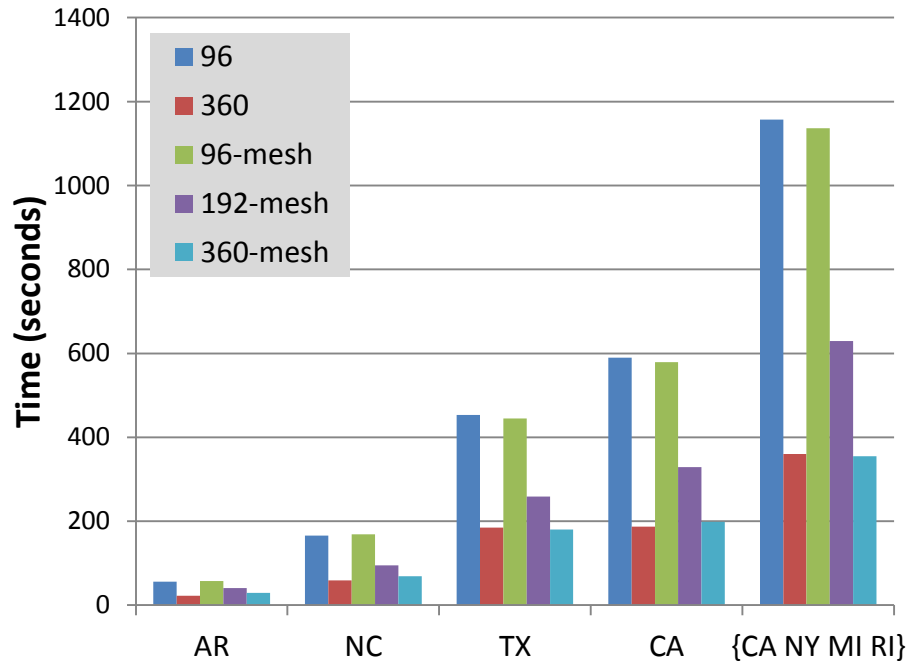
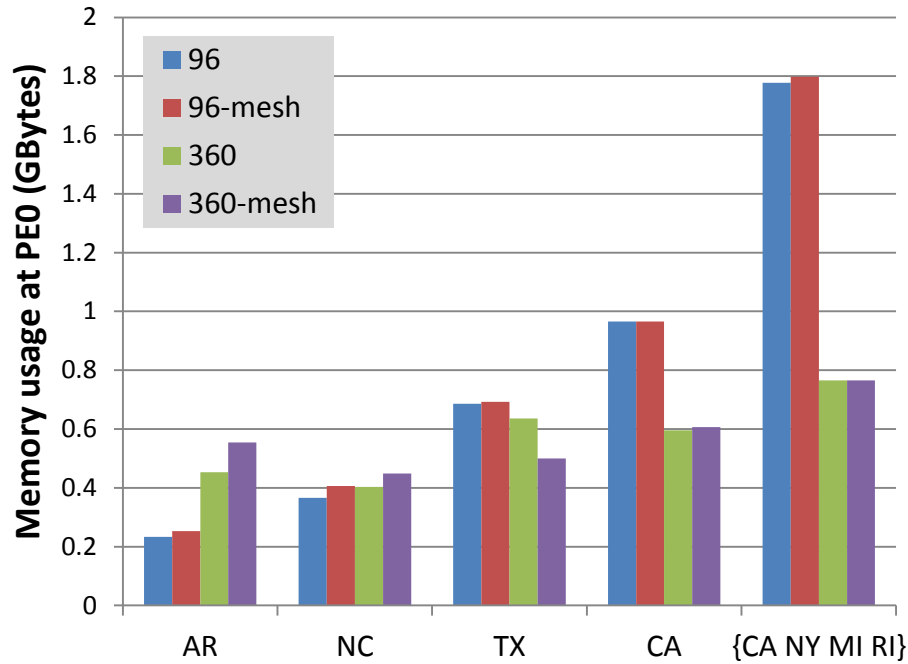
Calculated buffer space requirement

Total space per PE (Mbytes)

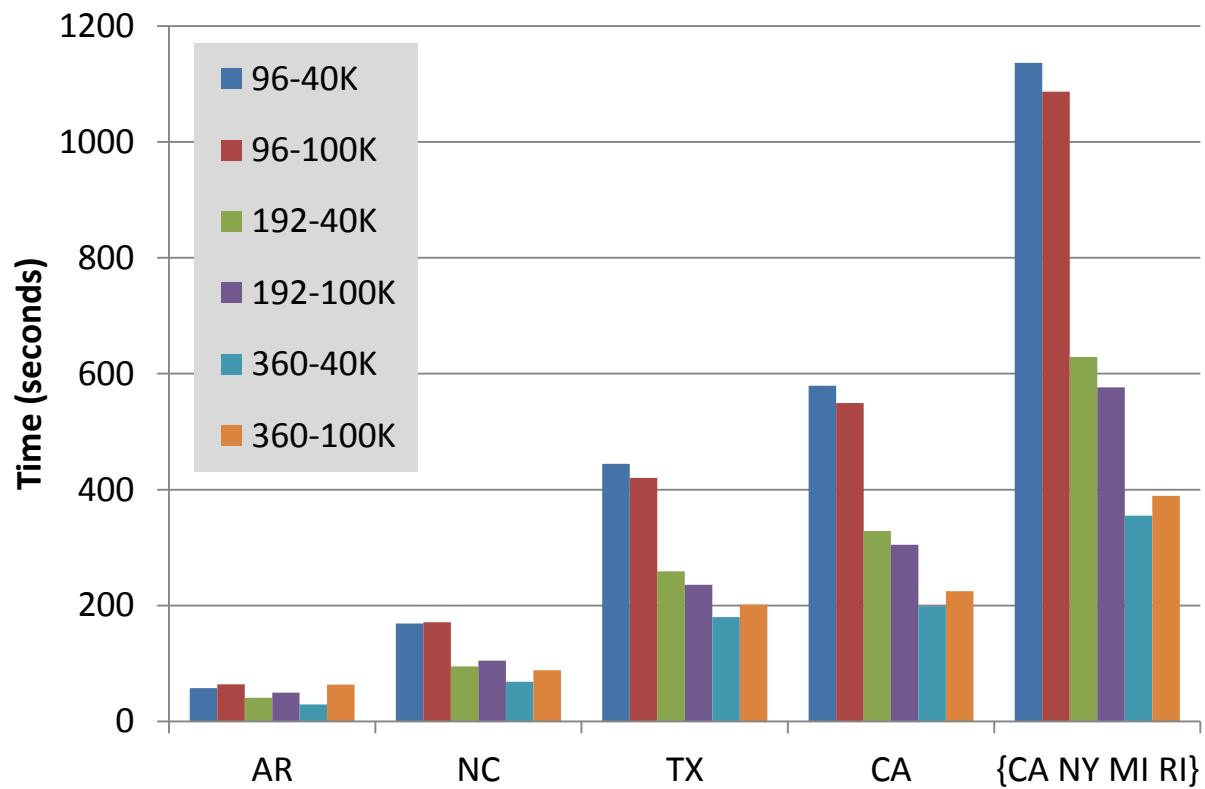
PE	mesh	manual
96	7.5	33.0
360	7.5	123.6

Space per destination (Kbytes)

PE	mesh	manual
96	80.0	351.6
360	21.3	351.6



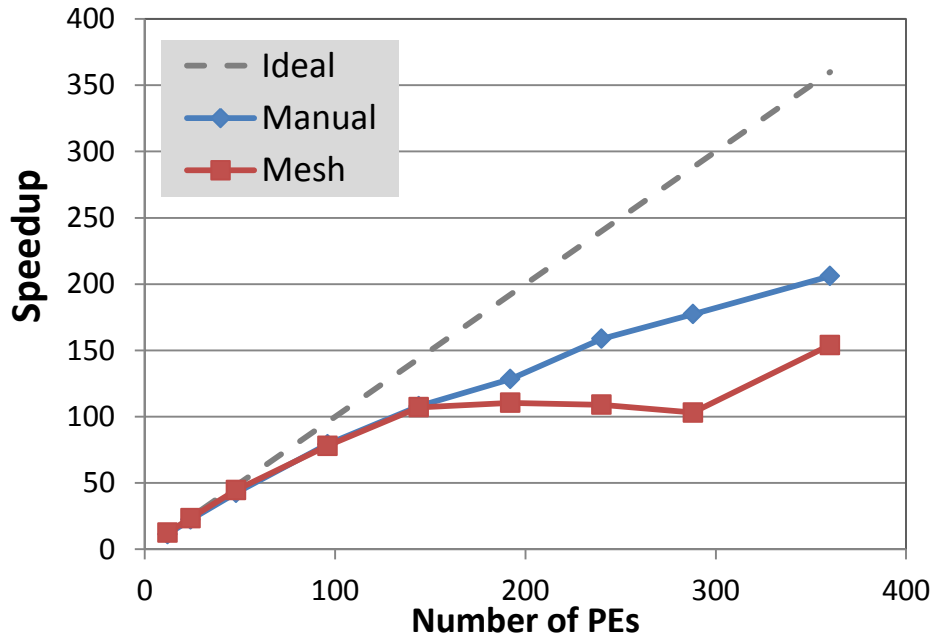
- Mesh-streaming : 1D topo / **100K threshold** / 160K capacity /48 bytes buffer unit
- Manual buffering : 10,000 messages per buffer / 36 bytes buffer unit
- 1 Chare per PE, 12 PEs per node
- Data population
 - AR (2,605,954) , NC (7,812,713) , TX (20,345,848), CA (33,153,148), {CA NY MI RI} (61,846,372)



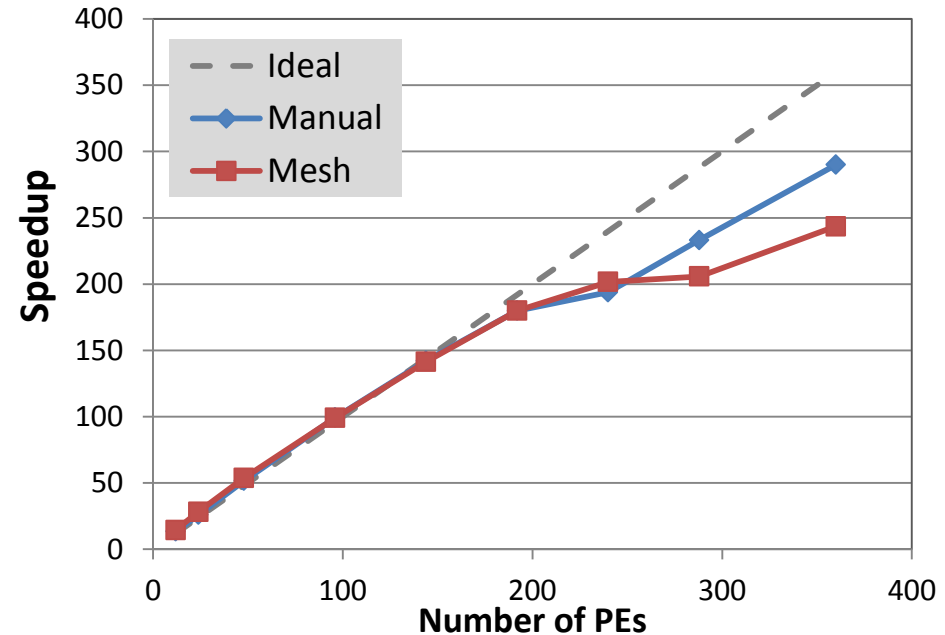
40K threshold vs. 100K threshold

Strong scaling with Mesh-streamer

AR (2,605,954)

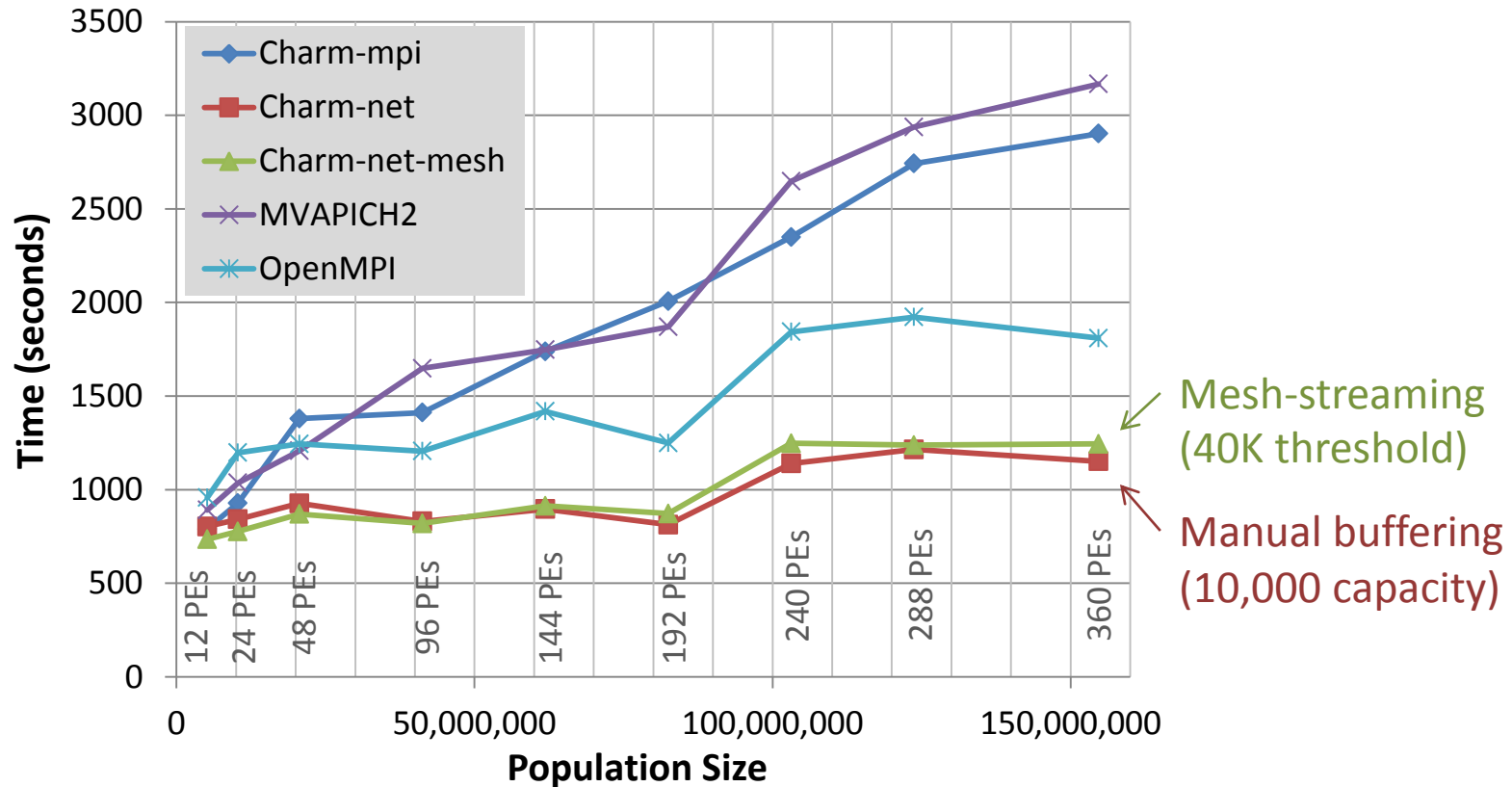


NC (7,812,713)



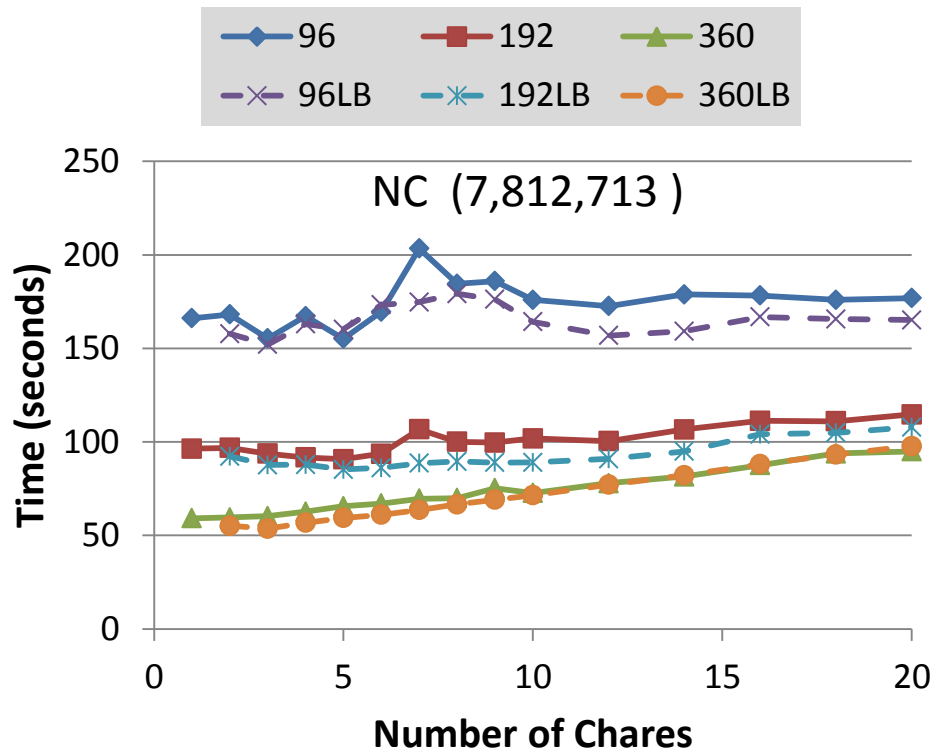
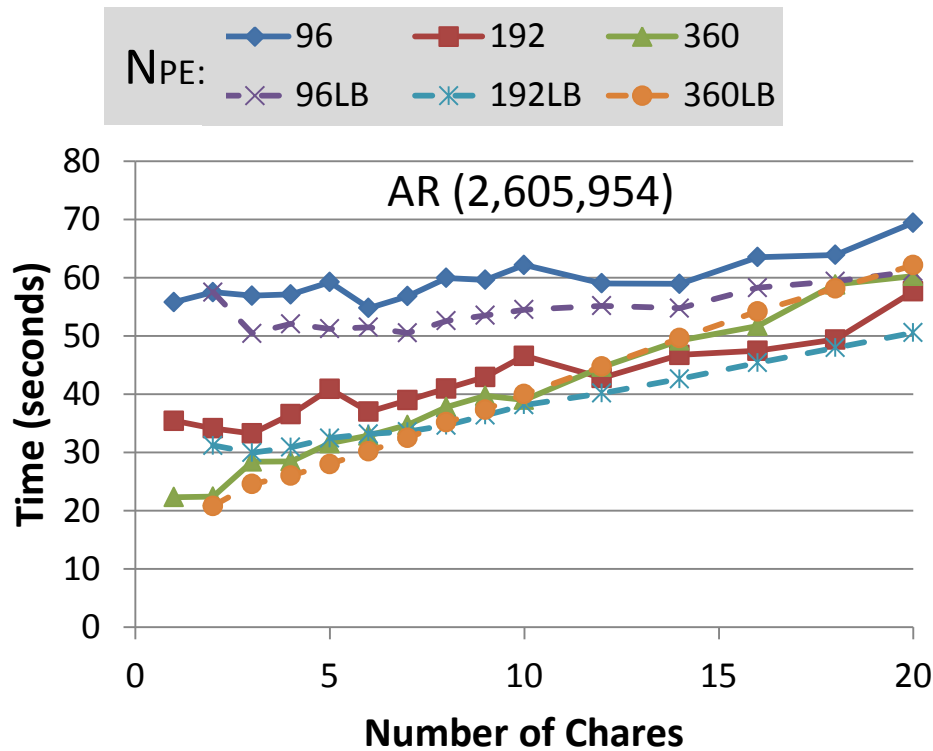
- Manual grouping buffer capacity: *10,000* messages per destination
- Mesh-streamer: 1-D topology, buffering threshold set to *40K* messages
- Both using Charm-net (ibverbs) on Shadowfax cluster
- Speedup is based on perf. of MPI-EpiSimdemics (MVAPICH2) using 1 core

Weak scaling with Mesh-streamer



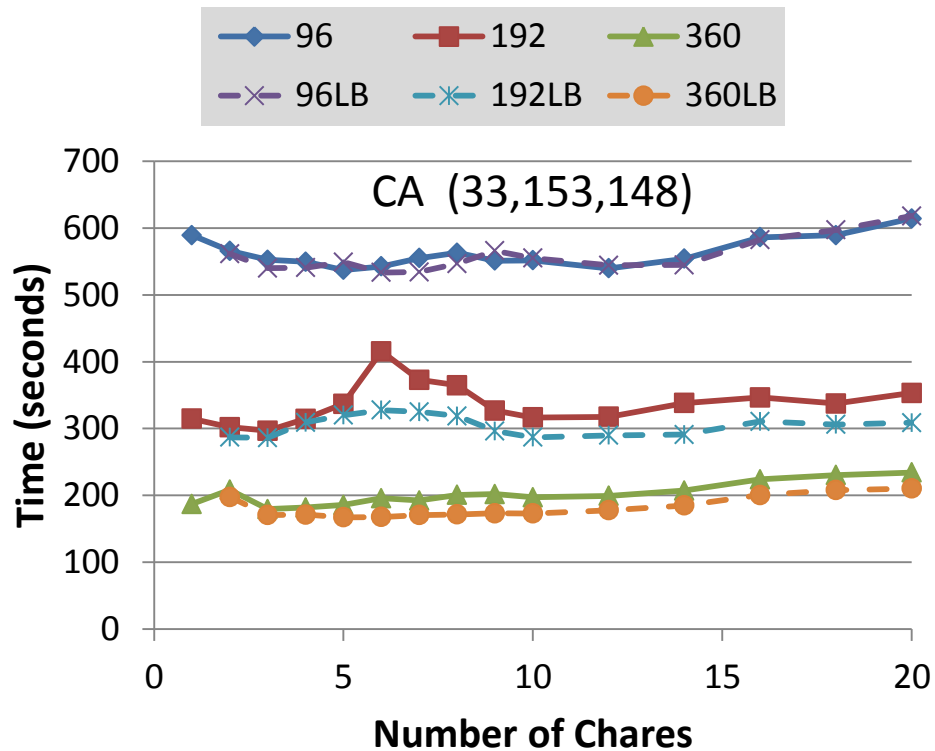
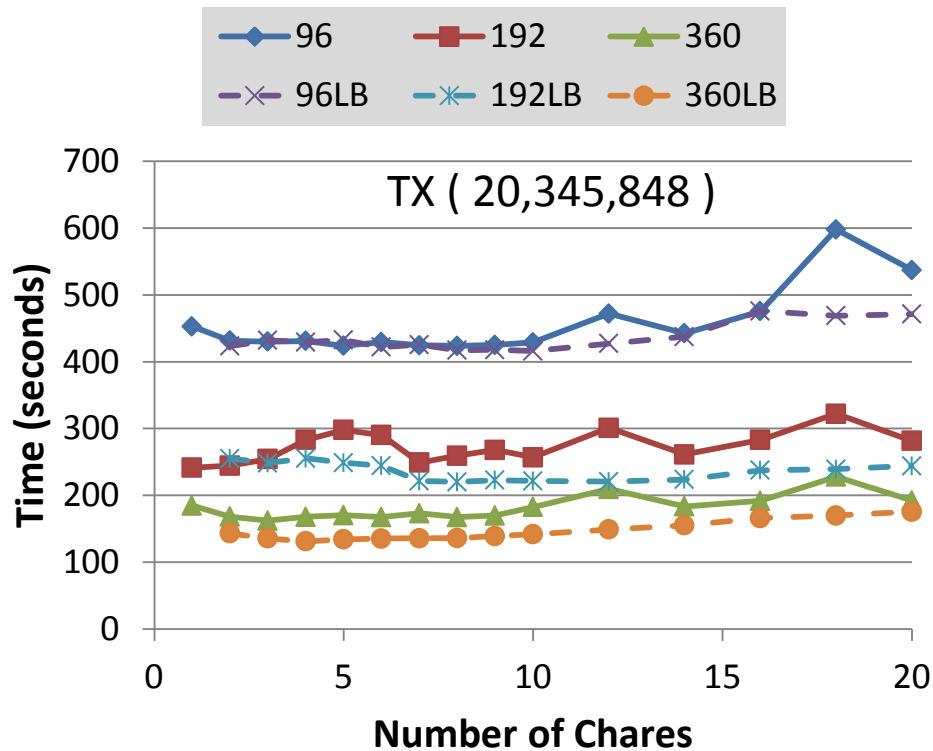
- Charm-net(-mesh) use net-linux-x86_64 (ibverbs)
- Charm-mpi uses mpi-linux-x86_64 (MVAPICH2)
- All tested on Shadowfax cluster (using 12 cores per node)

Multiple location Chares per PE and load balancing



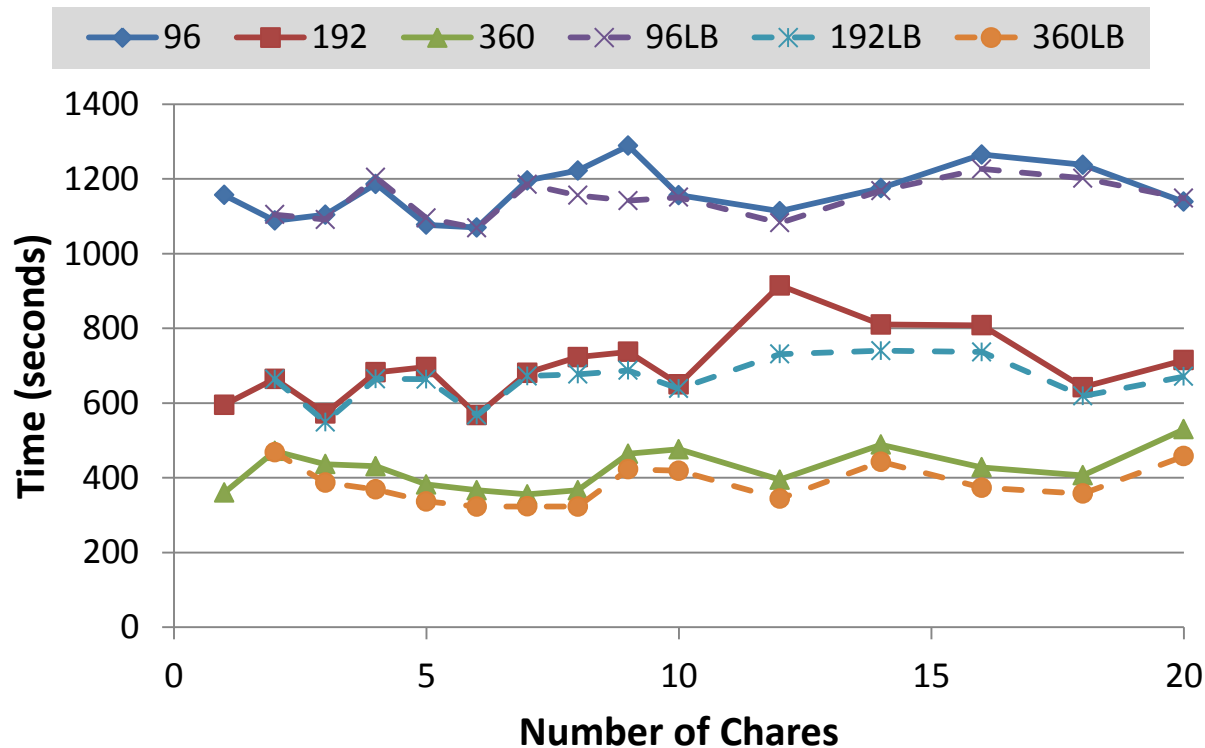
- *Solid line*: multiple location Chares per PE without load balancing
- *Dashed line*: **GreedyLB** once at the beginning and **RefineLB** at every 10th iteration
- Using Charm-net (ibverbs) with manual buffering (grouping 10,000 messages)

Multiple location Chares per PE and load balancing

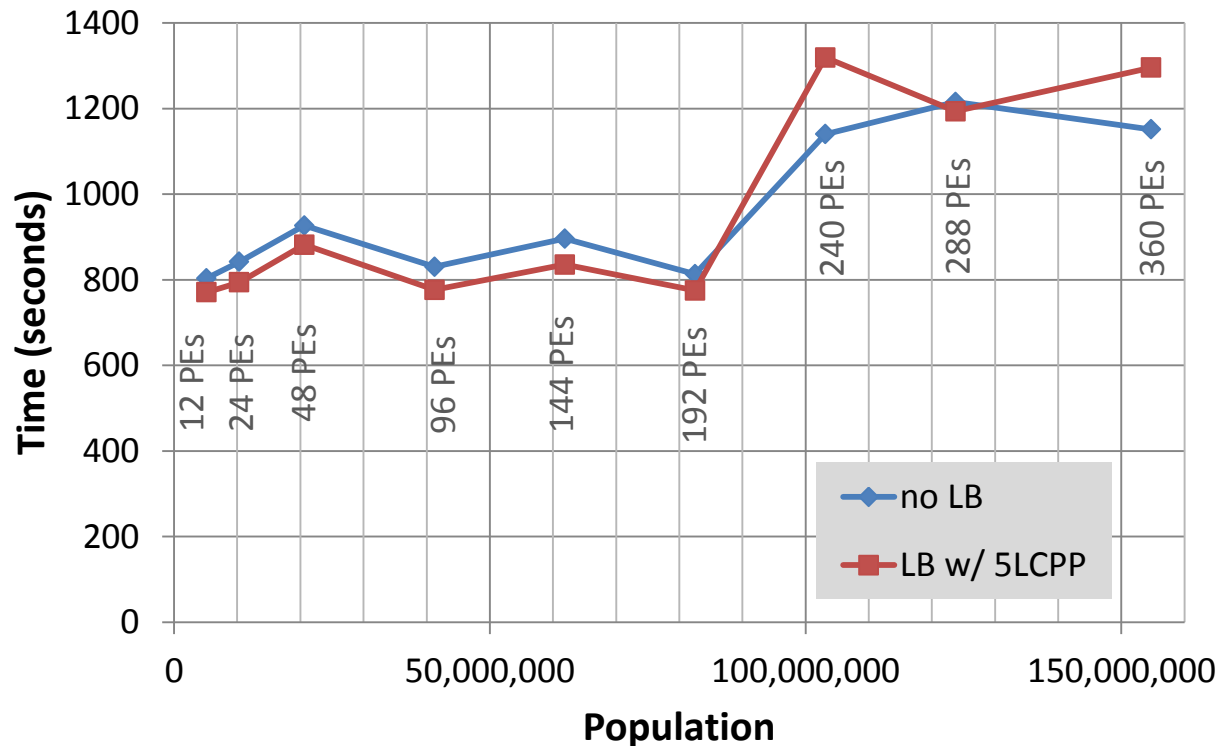


Multiple location Chares per PE and load balancing

{CA NY MI RI} (61,846,372)



Weak scaling with load balancing



- **GreedyLB** once at the beginning and **RefineLB** at every 10th iteration
- Location Chare as the migration unit
- Initially 5 location Chares per PE
- All tested on Shadowfax cluster (using 12 cores per node)
- Charm-net use net-linux-x86_64 (ibverbs)