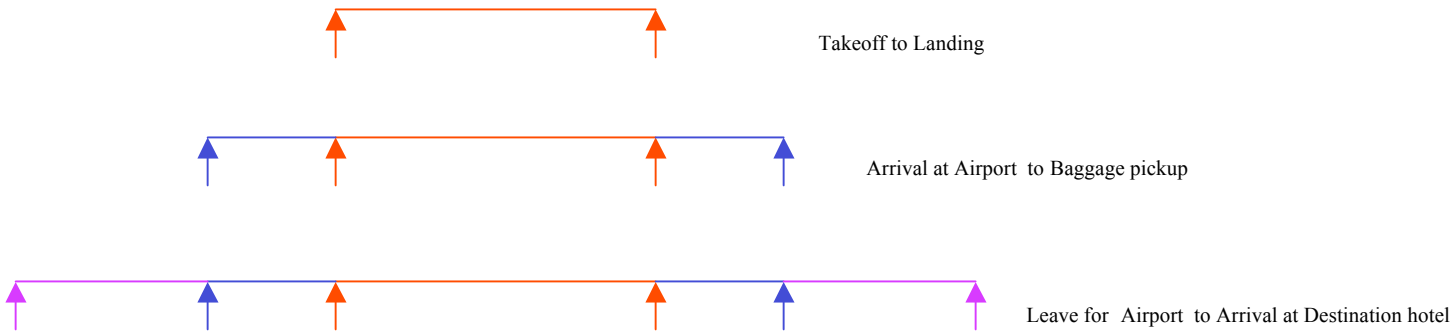
Latency and Throughput

- **Latency** (of task):
  - Time elapsed between start of the task and and its finish

Example: Travel from Houston to NY



Need to define start and end times by specifying the "task" carefully

Different subsystems may contribute to end-to-end latency

Each subsystem employs different latency reduction mechanisms

Latency and Throughput

- **Throughput** (of system or set of tasks):
  Rate of task completion: Number of tasks completed per unit time

  Instantaneous throughput (at time t):
  Rate of task completion in a "small" interval around time t
  Average throughput (in interval [0, T]):
  Number of tasks completed in interval / T

Latency and Throughput

Moving Large Data Sets to Data Center

Ship 1TB of data from Point of Creation to Point of Storage

FedEx:

Latency : 12 Hours

Throughput: 10^6 MB / (12 x 3600) sec ~ 23MB/sec

Not the usual understanding of throughput -- implicitly assumes small time granularity

Internet: 10Mb/sec link

Throughput = 10Mbps (assumes no errors/retransmissions, ignoring metadata overhead, link utilization)

Latency = 1TB/(10Mbps) = 800,000 sec ~ 224 hours ~ 10 days

With OC3 (155Mbits/sec) ~ 16 hours

Streaming Video

5 GB File 10Mbps link

Latency (for file download) = 5GB/10Mbps = 4000s ~ 1.1 hours (ideal)

Throughput: Smaller of 10MBps or Playback Rate

Latency (till start of transmission) : ~ 1s Latency (per packet) ~ ms

Latency and Throughput

- Travelers concern:

  Latency: Time for flight from Houston to NY --- 3 hours

- Airlines concern:

  Throughput: Rate at which it can move people

  - Assume 1 airplane of capacity 150 used on route
  - Maximum throughput: 150 persons / 6 hours = 25 pph

Latency and Throughput

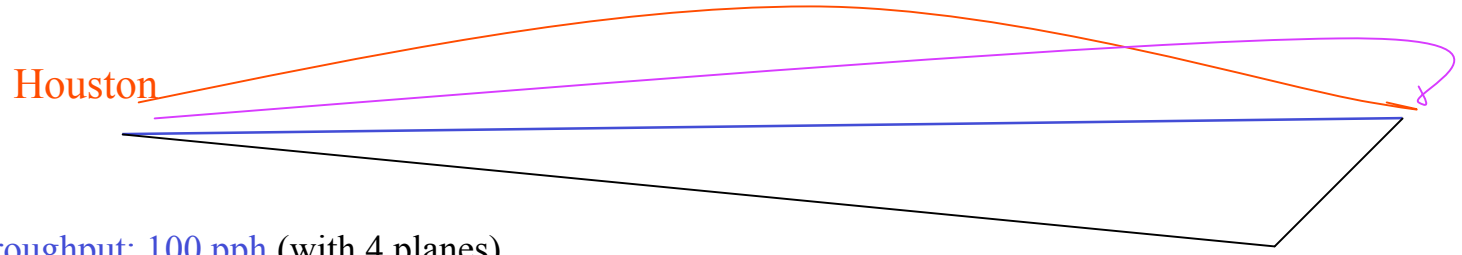
Improve Latency?

- Better Technology: Faster plane (materials, fuels, engine, …)
- Better Implementation: Faster route (wind pattern, distance, ….)

  - These will also usually improve the throughput
  - What if faster but smaller plane?
  - What if not enough passengers?

Improve Throughput?

- Bigger Plane: Capacity of 300 --- Throughput = 50 pph
- Multiple Planes:
  - Fly multiple independent routes simultaneously



Throughput: 100 pph (with 4 planes)

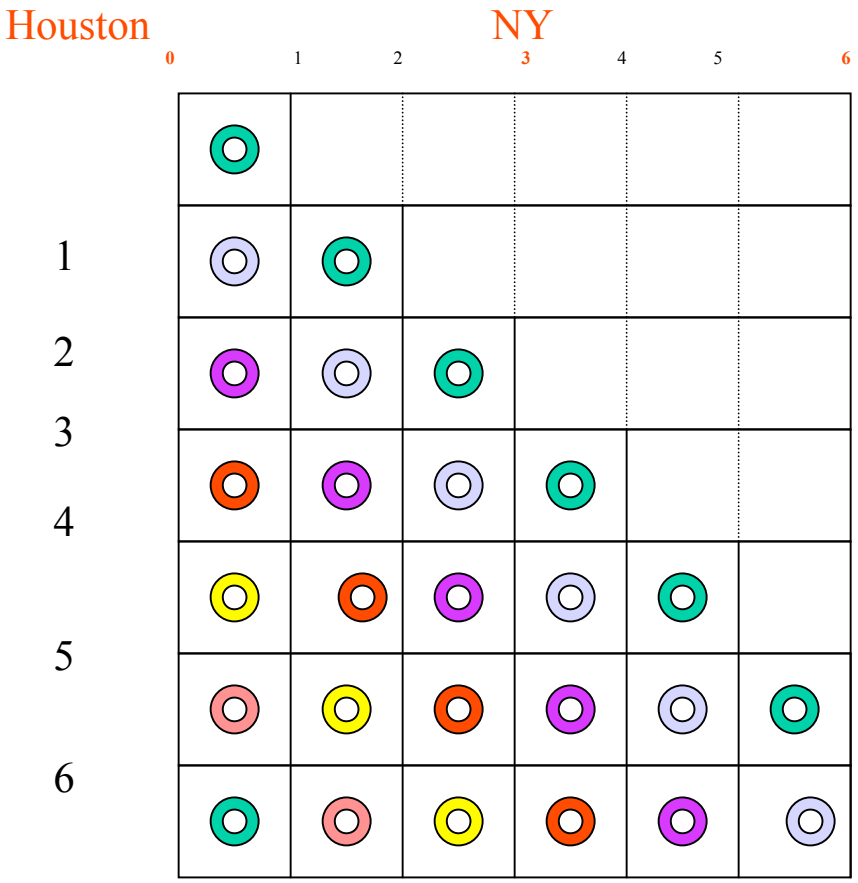
Theoretically continue to increase throughput by adding more planes/routes

Practical limits: Share segments and synchronize at intermediate stops (increase latency)

Popular and unpopular routes/aircraft raises load balancing issues

Latency and Throughput

- Pipeline along same route



Assuming 1 hour mandatory separation between successive flights on the same route

Can have 6 flights simultaneously in progress

Throughput: 150 persons / 1hr = 150 pph

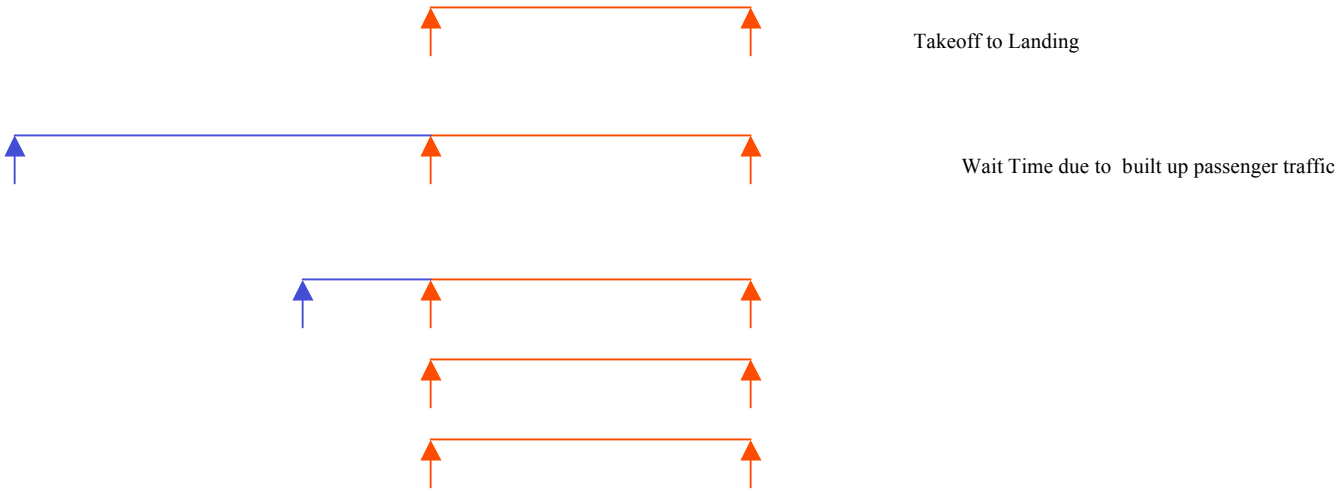
Note:

- Cannot use more planes (increase throughput) with given separation constraint
- Fundamental limits on minimum separation (*e.g.* one plane length!!!)

Latency and Throughput

- **Latency** (of task):
  - Time elapsed between start of the task and and its finish

Example: Travel from Houston to NY following storm



Can increasing throughput reduce latency?

Latency dependence on Load

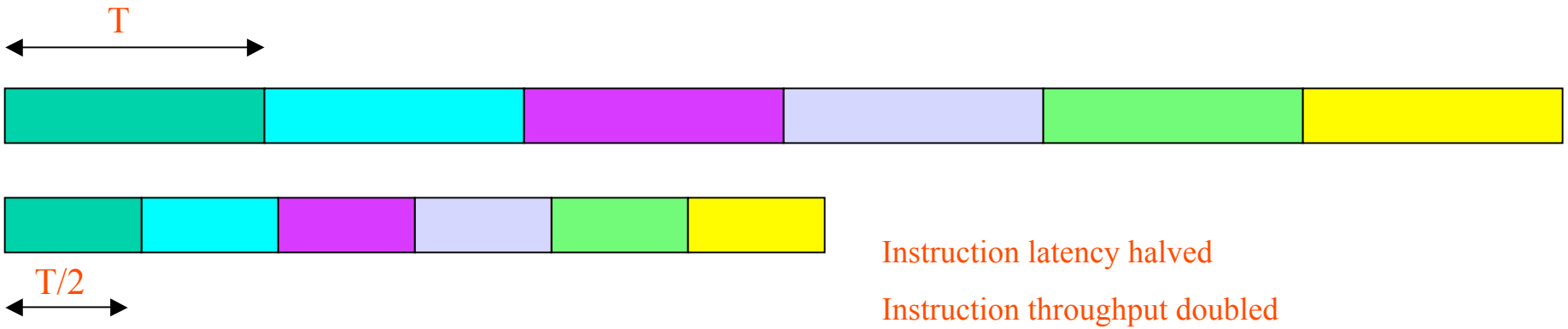
No load (lightly Loaded): No effect
Heavy Load: Reduces Queuing (wait) time

Latency and Throughput

Improve Latency?

1. Better Technology: Faster plane — Increase Clock Rate
2. Better Implementation: Faster route — Multi-cycle implementation
   - These will also improve the throughput

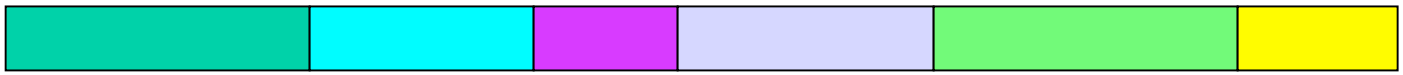
1.



Instruction latency halved

Instruction throughput doubled

2. Instructions have different latencies ---- average latency reduces depending on instruction mix



Average Latency = $\sum t_i w_i$

t_i : Latency for class i instructions

w_i : execution frequency of class i instructions

Average Throughput: $1 / \sum w_i t_i$

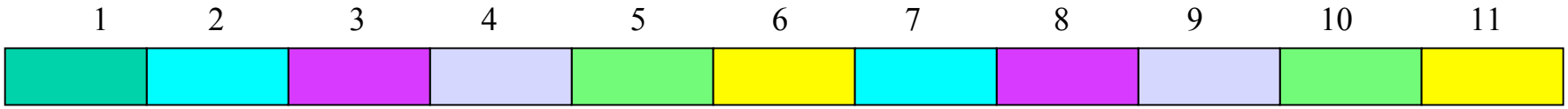
Best and worst case measures use pathological instruction mix

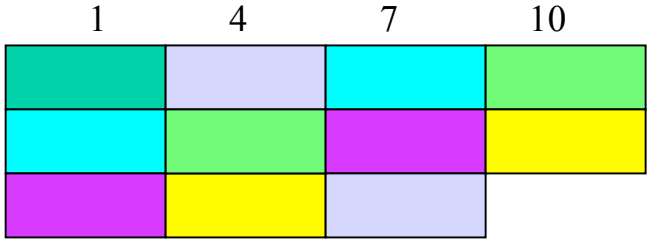
Latency and Throughput

Improve Throughput?

- Bigger Plane

Multiple Issue Processors
(VLIW, Superscalar, SMT)





T: 1 2 3 4 5 6 7 8 9 10 11

Each instruction bundle holds several instructions of the single-issue processor

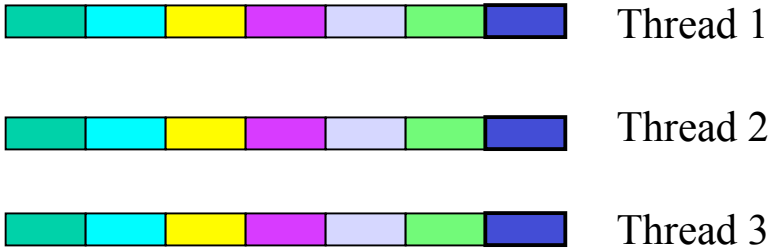
Throughput (potentially) increases by factor of bundle size

Latency and Throughput

Improve Throughput?

- Multiple Planes:
  - Fly multiple independent routes simultaneously
    - Multithreaded processors
    - Multi-core processors
    - Multiprocessors

Single-threaded loop execution



Multi-threaded loop execution with independent iterations

Latency and Throughput

Improve Throughput?

- Multiple Planes:
  - Pipeline along same route

Begin next instruction as soon as the previous instructions clears the first stage (path segment)

Overlapped pipelined execution

Latency and Throughput

Improve Throughput?

- Multiple Planes:
  - Pipeline along same route

Begin next instruction as soon as the previous instructions clears the first stage (path segment)

Overlapped pipelined execution