

Periodic Capacity Management under a Lead Time Performance Constraint

N.C. Buyukkaramikli^{1,2}

J.W.M. Bertrand¹

H.P.G. van Ooijen¹

1- TU/e IE&IS 2- EURANDOM

INTRODUCTION

- Using Lead time to attract customers is common
- Production systems target short & reliable lead times to be more competitive
- Lead Time Performance Constraint (LTPC):
 - E.g. Guarantee the job completion within **a week** with **95%** probability
 - A more ambitious LTPC has either a shorter lead time or a higher delivery performance target
- More Ambitious LTPC → More Capacity
- More Capacity → Higher Operating Costs
- **Capacity Flexibility** Can Soothe the Increase in Operational Costs.
 - Staffing, working overtime, etc...

INTRODUCTION

- Flexible Capacity Practices in real life are **periodic!**
 - External capacity pool → available at specific times
 - Working over/under time → periodic basis
 - Modus Operandi of most resource planning software
- Shorter Period Length → More Frequent Capacity Decisions
 - Better Tailoring of the Capacity
 - Frequency of capacity decision points → cost consequences
 - Example: overtime/under time decisions:
 - Company A(Weekly): A Decent Enough Time to Plan for Non-working time
 - Company B(At the Start of Each Day): No Private Life!
- We analyze Possible Benefits to be gained from Periodic Capacity Flexibility under a LTFC

OUTLINE

1. Literature
2. System Under Study
 - a. Capacity Related Costs
 - b. Problem Formulation
3. Analysis
4. Computational Study
5. Conclusion

LITERATURE REVIEW

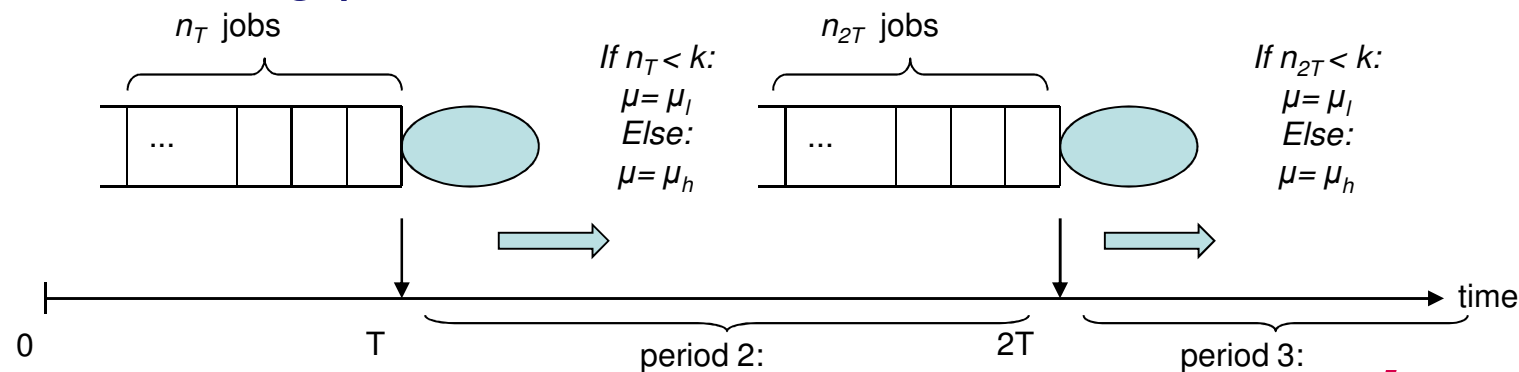
- Chenery (1952), Manne (1961)
 - Stylized Models, Holistic view of the effects of the capacity
- Holt et al. (1960):
 - Aggregate Planning over a horizon (No uncertainty)
- Pinker (1996):
 - DP Models (uncertainty in demand, not in service time)
- Literature on Queuing Control (e.g. Crabill (1972)):
 - Mostly Event Based Control Policies (Continuous-time)
 - Periodic Capacity Control → (transient probabilities)
 - Mostly Average Waiting (or Sojourn) Time is Penalized
 - Other Metrics (Variance of Sojourn Time or Lead Time Performance) → sojourn time distribution
- **Our Study: Modeling the Periodic Capacity Control under a LTPC (in a Queuing Context)**

SYSTEM UNDER STUDY

- A Single Production System
- Arrivals of the jobs follow a unit Poisson process
- Time requirement of each job: exponentially distributed.
- Jobs are being served based on a FCFS rule.
- The system operates with a LTPC that guarantees a ratio (γ) of timely deliveries within a fixed lead time L .
- Capacity corresponds to the service rate of the system.

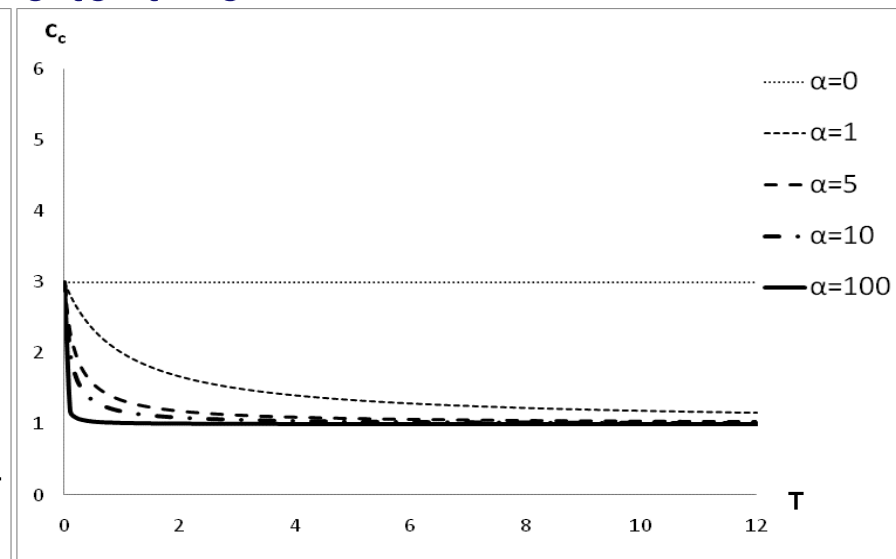
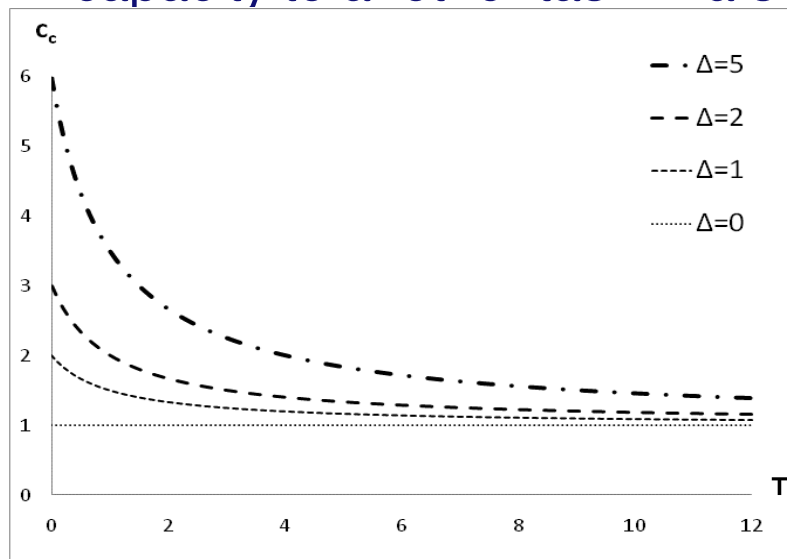
MODEL: Periodic Capacity Policy

- We assume a workload dependent two-level threshold type of periodic capacity policy!
- Every " T " time units the system is observed.
- Let $X(t)$ denotes the number of jobs at time t .
- The server rate is set to μ_l at the start of period n if: $X((n-1)T) < k$, otherwise it is set to μ_h for $n=1,2,\dots$ & $\mu_h \geq \mu_l$
- k : switching point



MODEL: Capacity Related Costs

- $\mu_l \rightarrow$ permanent capacity & $\mu_h - \mu_l \rightarrow$ contingent capacity
- Permanent Capacity Cost per unit time: c_p
- Contingent Capacity Cost per unit time: $c_c = c_p + \Delta / (1 + \alpha T)$
 - Δ : Lost opportunity cost of contingent capacity for being ready to be deployed at the start of each period.
 - If $T \uparrow$, contingent capacity is more easily assigned to another task
 - α : Mitigation factor of T . If $\alpha \uparrow$, easier to assign the contingent capacity to another task in a shorter time



MODEL: Problem Formulation

- The **A**verage **C**apacity **U**sage from given permanent and contingent capacity levels and switching point with a certain period length can be found.
 - Let $ACU(\pi(\mu_l, \mu_h, k, T))$ denote the **A**verage **C**apacity **U**sage
- Given c_p and c_c , **A**verage **C**apacity **C**osts can be derived from $ACU(\pi(\mu_l, \mu_h, k, T))$:
 - $ACC(\pi(\mu_l, \mu_h, k, T)) = c_p * \mu_l + (ACU(\pi(\mu_l, \mu_h, k, T)) - \mu_l) * c_c$
- Let $S(\pi(\mu_l, \mu_h, k, T))$ denote the resulting sojourn time
- The Optimization Problem can be formulated:

$$\min_{T, \mu_h, \mu_l, k} ACC(\pi(k, \mu_l, \mu_h, T)) \longrightarrow \text{From Analysis Part 1}$$

s.t.

$$P(S(\pi(k, \mu_l, \mu_h, T)) > L) \leq 1 - \gamma$$

From Analysis Part 2

ANALYSIS-I: Limiting Probabilities of the Number of Jobs

- Truncate infinite $M/M/1$ system to a $M/M/1/K$ system with big enough finite waiting room K .
- **Property 1:** The number of jobs at the start of each period ($X((n-1)T)$) satisfies the Markov property for $n=1,2,\dots$
- $Q(T)$ is transition matrix of the DTMC: $X(nT)$ for $n=0,1,\dots$
 - $Q_{ij}(T)$ can be derived from transient analysis formulas.
 - These formulas involve infinite sum of Bessel Functions...
 - We develop a numerical method based on eigen-decomposition.

ANALYSIS-I: Limiting Probabilities of the Number of Jobs

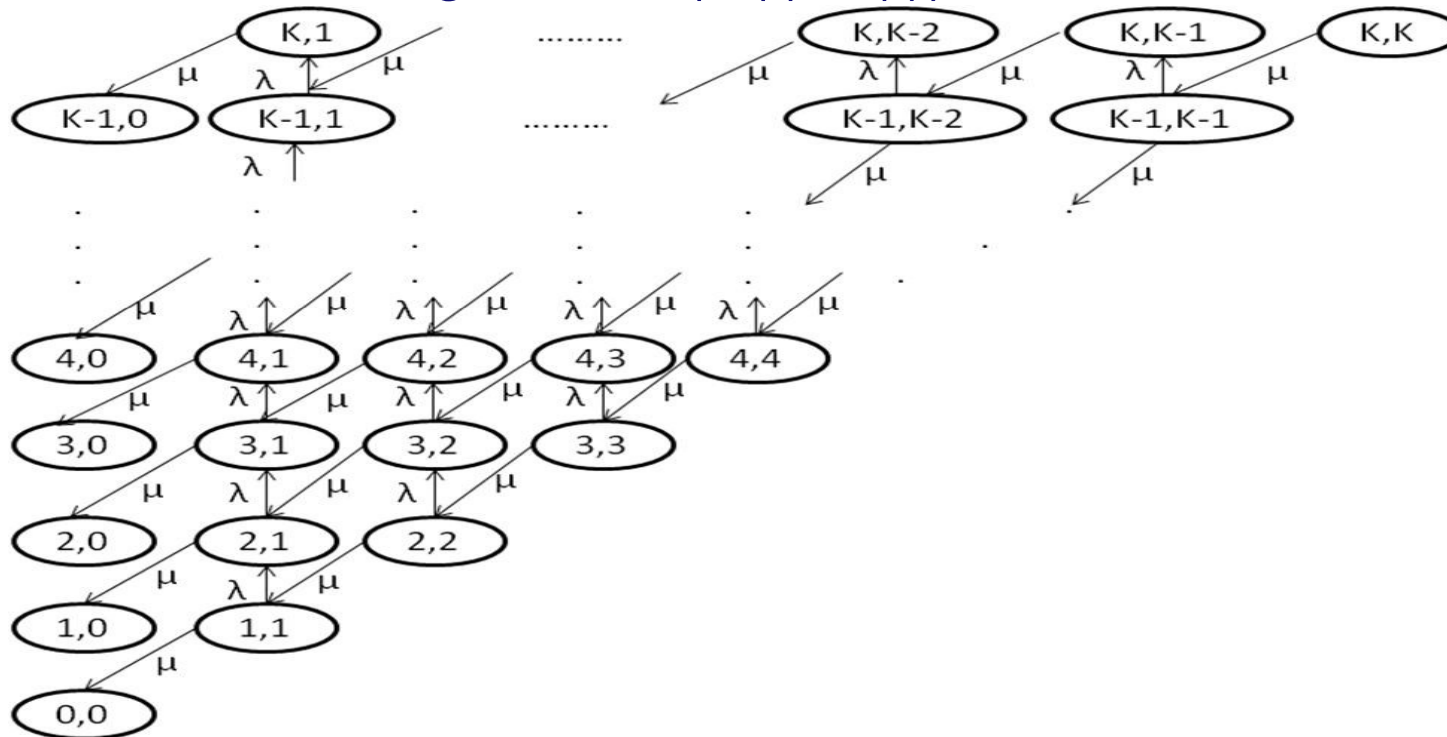
- From $Q(T)$, we can obtain $v(T)$, the steady state probability vector of the number of jobs at the end of each period.
- $P(n \text{ jobs @ the start of a period}) = v_n(T)$.
- $ACU(\pi(\mu_l, \mu_h, k, T)) = \sum_{i=0}^{k-1} \mu_l \times v_i(T, \pi) + \sum_{i=k}^K \mu_h \times v_i(T, \pi)$

ANALYSIS-II: Sojourn Time Distribution

- Extended stochastic process $(X(t), Y(t))$.
 - $X(t) \rightarrow$ number of jobs in the system
 - $Y(t) \rightarrow$ the position of a tagged customer in the queue.
- $0 \leq X(t) \leq Y(t) \leq K$ and $Y(t_2) \leq Y(t_1)$ for $t_1 \leq t_2$
- When a tagged job finds $m-1$ jobs in the queue at its arrival at time t , $(X(t), Y(t)) = (m, m)$ for $0 < m \leq K$.
- When $Y(t)=0$, the tagged job's service is complete.
(Absorbing States)

ANALYSIS-II: Sojourn Time Distribution

- Transition diagram of $(X(t), Y(t))$:



- When $\mu_l = \mu_h = \mu$

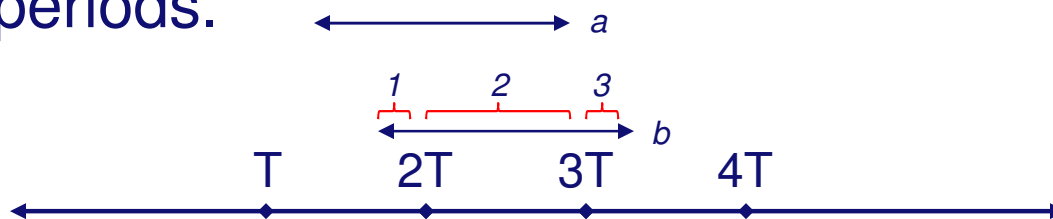
ANALYSIS-II: Sojourn Time Distribution

- **Property 2:** Under a two-level capacity policy, $(X(nT), Y(nT))$ has Markovian property, and $A(T)$ is the transition matrix of the $(X(nT), Y(nT))$ DTMC.
- We use $A(T)$ in the analysis of the sojourn time.
 - Construction of $A(T)$ requires the transient analysis of $(X(t), Y(t))$ with $\mu = \mu_h$ and $\mu = \mu_l$
 - Uniformization Technique for the transient analysis of the $(X(t), Y(t))$
- $P(S > L)$ = Probability that an arriving job is not absorbed after L units of time.

ANALYSIS-II: Sojourn Time Distribution

Initial state: Arrival time, number of jobs upon arrival and the capacity level of the system upon arrival

Example: Suppose $T < L \leq 2T$: L can spread over either 2 (a) or 3 (b) periods.



Given initial state upon arrival, keep track of $(X(t), Y(t))$ during:

1. The time between job arrival and end of the first period
 2. Periods between the first and the last period
 3. The last period
- After unconditioning on the initial states
 - $P(S > L)$ can be derived.

ANALYSIS

- Randomized Action:
 - $k \rightarrow$ not necessarily an integer: if k customers: μ_l with p , μ_h with $1-p$
 - Easy to incorporate to the analysis

$$ACU(\pi(\mu_l, \mu_h, k, T)) = \mu_l + (\mu_h - \mu_l) \left(\sum_{i=k+1}^K v_i(T, \pi) + (1-p) \times v_k(T, \pi) \right)$$

- Given a LTPC with lead time L and with γ performance and c_p and c_c cost structure, following decisions should be taken:
 1. What should be the size of period length T ?
 2. What should be the size of permanent & contingent capacity?
 3. At which workload levels should we use the contingent capacity?

COMPUTATIONAL STUDY

- Scaling arrival rate and permanent capacity cost rate to 1
- 3 Lead Time Performance Constraints:
 - Low-Ambition LTPC: $P(S > 10) < 0.90$
 - Ambition LTPC: $P(S > 5) < 0.90$
 - High-Ambition LTPC: $P(S > 5) < 0.95$
- An Important Reference Point:
 - $\mu_{L,\gamma}$: Min. constant service rate in a $M/M/1$ queue to satisfy the LTPC with lead time L and with γ performance.
 - $\mu_{L,\gamma}$ increases with the ambition level of the LTPC

COMPUTATIONAL STUDY

Search algorithm

- For every LTPC, create the Ω_l and Ω_h sets
 - $\Omega_l = \{\mu_{l1}, \mu_{l2}, \mu_{l3}, \mu_{l4}, \mu_{l5}\}$ and $\Omega_h = \{\mu_{h1}, \mu_{h2}, \mu_{h3}, \mu_{h4}, \mu_{h5}\}$
 - $\mu_{li} = (i/6) * \mu_{L,\gamma}$ and $\mu_{hi} = \mu_{li} + \mu_{L,\gamma}$
- For every candidate period length T , run the search algorithm

For $i=1$ to 5 choose
 For $j=1$ to 5 choose
 $k=0$;
 Do: $k=k+0.1$;
 $p=k - \lceil k \rceil$;
 while $P(S>L) \leq 1-\gamma$ under $\pi(\mu_{li}, \mu_{hj}, k, T)$ with p ;
 $k^*=k$;
 $p^* = k^* - \lceil k^* \rceil$;

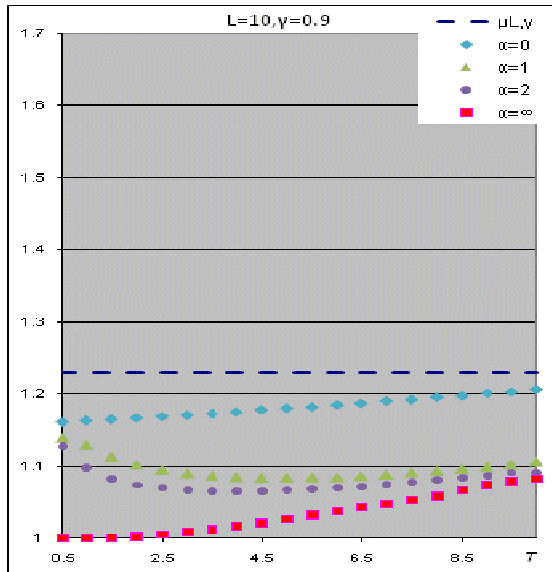
$ACU(\pi(\mu_{li}, \mu_{hj}, k^*, T)) = ACU(\pi(\mu_{li}, \mu_{hj}, \lceil k^* \rceil, T))$ with p^* ;

Calculate $ACU(\pi(\mu_{li}, \mu_{hj}, k^*, T))$;
 Find $ACC(\pi(\mu_{li}, \mu_{hj}, k^*, T)) = c_p + (ACU(\pi(\mu_{li}, \mu_{hj}, k^*, T)) - \mu_{li}) * c_c$;
 End For
 End For

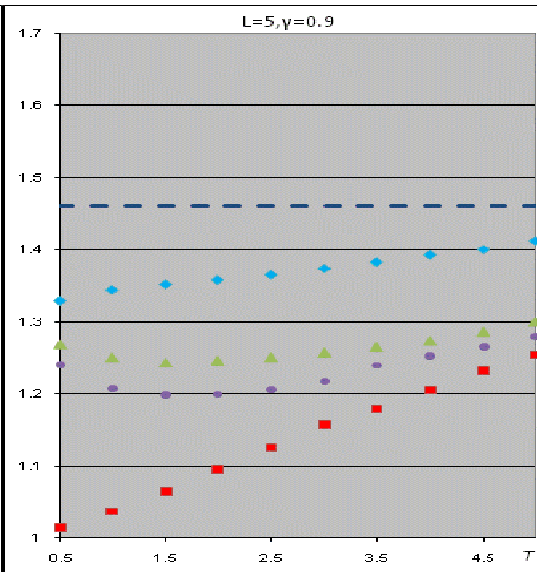
$ACC^*(T) = \min_{i,j} \{ ACC(\pi(\mu_{li}, \mu_{hj}, k^*, T)) \}$;

COMPUTATIONAL STUDY

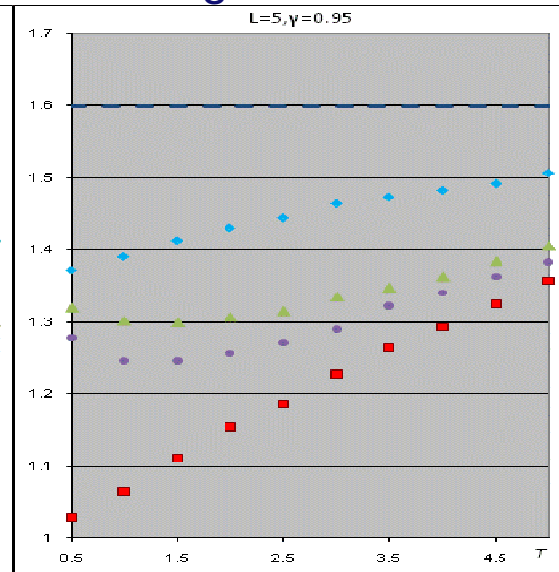
Low-Ambition



Ambition



High-Ambition



$ACC^*(T)$ for increasing T under three LTPC for $c_p=1$, $\Delta=1$ and $\alpha=0,1,2,\infty$

- Best period Length T^* is not necessarily the minimum possible T

Savings of Capacity Flexibility under T^* compared to $\mu_{L,\gamma}$

	$\Delta=0, \alpha=1$	$\Delta=0.5, \alpha=1$	$\Delta=1, \alpha=1$
$L=10, \gamma=0.9$	18.72%	13.74%	11.92%
$L=5, \gamma=0.9$	30.55%	19.36%	14.81%
$L=5, \gamma=0.95$	35.68%	23.78%	18.70%

	$\Delta=1, \alpha=0$	$\Delta=1, \alpha=1$	$\Delta=1, \alpha=2$
$L=10, \gamma=0.9$	5.60%	11.92%	13.40%
$L=5, \gamma=0.9$	9.02%	14.81%	17.95%
$L=5, \gamma=0.95$	14.23%	18.70%	22.05%

- The system is more insensitive with regard to changes in LTPC!

CONCLUSION & FUTURE RESEARCH

- Periodic Capacity Management under LTPC in a queuing framework that incorporates the transient effects
- Modeling Framework & Analysis is extendable.
 - Policies that update both the number of servers and service rate jointly.
 - DP problem framework
 - No LTPC but penalizing the average sojourn(or waiting) time
- Managerial insights on the effect of period length T in a periodic capacity management problem.
- Future Research Question: How to integrate the consequences of the capacity decisions to the decision making of the customers on the lead time performance constraint?



**THANK YOU FOR YOUR
ATTENTION**