دانشکده مهندسی کامپیوتر

عنوان درس:

ارزیابی کارایی سیستم‌های کامپیوتری

Performance Evaluation of Computer Systems (PECS)

جلسه ۱۳: اصول اولیه مدلهای صف

مدرس:

محمد عبداللهی ازگمی

(Mohammad Abdollahi Azgomi)

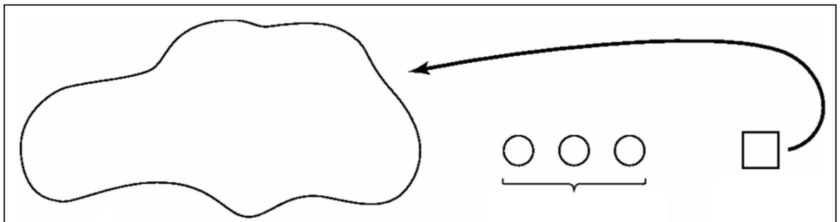
azgomi@iust.ac.ir

فهرست مطالب

- مقدمه
- ویژگیهای سیستم‌های صف
- نمادهای مدل‌های صف
- معیارهای کارایی سیستم‌های صف
- رفتار حالت پایدار مدل‌های صف مارکوفی
- کاربردهای مدلهای صف در ارزیابی کارایی سیستم‌های کامپیوتری

مقدمه

- **مدل‌های صف (queueing models)** در زمینه‌های متنوعی دارای کاربرد هستند.
- در شکل زیر نمای ساده‌ای از مدلهای صف ارائه شده است:



مقدمه

- **مدلهای صف** ابزار قدرتمندی را برای ارزیابی کارایی **سیستم‌های صف** (queueing systems) در زمینه‌ها و کاربردهای مختلف برای تحلیل‌گران سیستم فراهم می‌کنند.
- **معیارهای کارایی** مهمی که می‌توان با ارزیابی مدلهای صف بدست آورد عبارتند از:
  - بهره‌وری سرویس‌دهنده (server utilization)،
  - طول خط انتظار (length of waiting lines) و
  - تاخیر مشتریان (delay of customers).
- محاسبه معیارهای فوق برای سیستم‌های ساده با روشهای ریاضی و تحلیلی امکان‌پذیر است.
- اما برای مدلهای واقعی سیستم‌های پیچیده، شبیه‌سازی کامپیوتری مورد استفاده قرار می‌گیرد.
- در ادامه پس از آشنایی با ویژگیهای سیستم‌های صف، نمادها، مدلهای صف مارکوفی و کاربردهای این مدلها در ارزیابی کارایی معرفی می‌شوند.

ویژگیهای سیستم‌های صف

- **مشتری (customer):** به هر درخواست‌کننده که به یک سیستم (دستگاه، محل ارائه خدمت، سیستم‌کامپیوتری، و غیره) وارد شده و سرویس دریافت می‌کند **مشتری** آن سیستم گفته می‌شود.
  - **مثال:** مردم، ماشین‌ها، کامیون‌ها، پیغامهای پست الکترونیکی و غیره.
- **سرویس‌دهنده (خدمت‌دهنده) (server):** به هر **منبعی** (resource) که سرویس‌های درخواستی را به مشتریانش ارائه می‌کند.
  - **مثال:** تعمیرکاران، باندهای فرودگاه، پردازنده یک سیستم کامپیوتری، و غیره.

ویژگیهای سیستم‌های صف

- **جمعیت (population):** به مشتریان بالقوه یک سیستم صف گفته می‌شود که ممکن است وارد سیستم شده و از سرویس‌های آن استفاده کنند.
- جمعیت صف می‌تواند **متناهی** (finite) یا **نامتناهی** (infinite) باشد:
  - **مدل‌های دارای جمعیت متناهی** (finite population model): اگر نرخ ورود (arrival rate) مشتریان به سیستم صف وابسته به تعداد مشتریان در حال سرویس یا منتظر دریافت سرویس باشد، مدل دارای جمعیت متناهی است.
  - **مدل‌های دارای جمعیت نامتناهی** (infinite population model): اگر نرخ ورود (arrival rate) مشتریان به سیستم صف متاثر از تعداد مشتریان در حال سرویس یا منتظر دریافت سرویس نباشد، مدل دارای جمعیت نامتناهی است. سیستم‌هایی که دارای جمعیت بزرگی از مشتریان بالقوه باشند، از این نوع هستند. مثلاً همه ما جزء مشتریان بالقوه بانک هستیم. بنا برا این بانک دارای جمعیت نامتناهی است.

ویژگیهای سیستم‌های صف (ادامه)

- **ظرفیت سیستم (system capacity):** محدودیتی را که برای تعداد مشتریان منتظر یا در حال سرویس یک سیستم وجود دارد را مشخص می‌کند:
  - **ظرفیت محدود (limited capacity):** برای مثال یک کارواش اتوماتیک ممکن است که دارای ظرفیت ۱۰ برای ماشینهای منتظر باشد. تا یکی از ماشین‌های قبلی شسته و خارج نشود، ماشین جدید پذیرش نمی‌شود.
  - **ظرفیت نامحدود (unlimited capacity):** برای مثال صف فروش بلیط کنسرت می‌تواند نامحدود باشد و به همه کسانی که منتظرند بلیط فروخته شود.
- **اغلب اوقات ظرفیت سیستم محدود است.** یعنی اگر تعداد مشتریان وارد شده به سیستم از حدی بیشتر شود دیگر به مشتریان جدید اجازه ورود به سیستم داده نمی‌شود و قبل از ورود **طرد** (reject) می‌شوند.
- ظرفیت صف (queue capacity) **هم مطرح است که شامل تعداد مکانهای صف انتظار است.**
- **اما ظرفیت سیستم صف، شامل ظرفیت صف به‌علاوه تعداد مشتریان در حال سرویس (به تعداد سرویس‌دهنده‌های موازی) است.**

ویژگیهای سیستم‌های صف (ادامه)

- **فرآیند ورود (arrival process):** بر حسب زمانهای بین ورود (interarrivals) مشتریان پی‌درپی مشخص می‌شود.
- **زمانهای بین ورود می‌تواند تصادفی یا زمانبندی شده باشد:**
  - **ورودهای تصادفی (random arrivals): اغلب زمانهای بین ورود به‌وسیله توزیع‌های احتمالی مشخص می‌شود.**
    - اغلب از **توزیع‌های یکسان و مستقل** (iid: independent and identical distributions) استفاده می‌شود.
    - در حالت کلی ورودها ممکن است به‌صورت توده‌ای یا دسته‌ای باشد (bulk or batch arrivals) یا به‌صورت مرتبط‌به‌هم (correlated arrivals) بوده و iid نباشند.
    - مهمترین مدل، فرآیند ورود پواسان (با نرخ λ) است، که در آن A_n زمان بین ورود مشتری (n-1)–ام و n–ام را مشخص می‌کند و یک توزیع نمایی (با میانگین $1/\lambda$) است.
  - **ورودهای زمانبندی شده** (scheduled arrivals): زمانهای بین ورود می‌تواند ثابت یا ثابت به‌اضافه یا منهای یک مقدار تصادفی برای نشان دادن دیر یا زود شدن زمانهای بین ورود باشد.
    - برای مثال بیماران یک پزشک یا پروازهای ورودی به یک فرودگاه ممکن است دیر یا زود وارد شوند.

ویژگیهای سیستم‌های صف (ادامه)

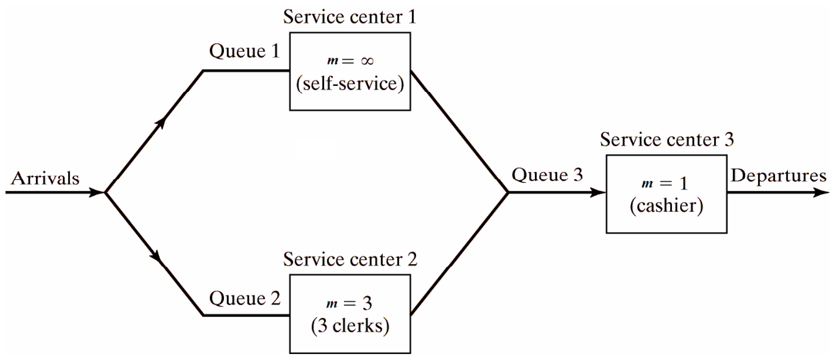
- **توزیع زمان سرویس (service time distribution):** یعنی بازه زمانی صرف‌شده برای سرویس مشتریان از چه مدلی تبعیت می‌کند؟
- زمانهای سرویس متناظر با مشتریان وارده شده پی‌درپی با $\mathbf{S_1, S_2, S_3, \ldots}$ نشان داده می‌شوند:
  - این زمانها ممکن است ثابت یا تصادفی باشند.
  - مجموعه $\mathbf{\{S_1, S_2, S_3, \ldots\}}$ اغلب با دنباله‌ای از متغیرهای تصادفی iid مشخص می‌شود. برای مثال، توزیع‌های نمایی، ویبول، گاما و لاگ‌نرمال برای این منظور استفاده می‌شوند.

ویژگیهای سیستم‌های صف (ادامه): مکانیسم سرویس

- **مکانیسم سرویس** (service mechanism): یکی دیگر از ویژگیهای سیستم‌های صف است.
- یک سیستم صف می‌تواند شامل تعدادی مراکز سرویس (service centers) و صف‌های به هم متصل (interconnected queues) باشد.
  - هر مرکز سرویس شامل تعدادی سرویس‌دهنده، m، که به‌طور موازی کار می‌کنند.
  - مشتری که در سر صف بوده و نوبت آن است توسط اولین سرویس‌دهنده در دسترس سرویس‌دهی می‌شود.

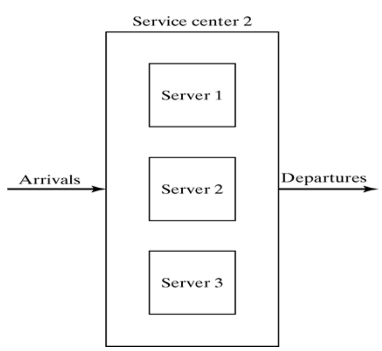
ویژگیهای سیستم‌های صف (ادامه): مکانیسم سرویس

- برای مثال، در یک حراجی که کالاها را با تخفیف می‌فروشد ممکن است که مشتریان مطابق شکل زیر خودشان کالاها را انتخاب و بردارند (self-service) (مرکز سرویس ۱) و سپس قیمت آنها را به صندوق‌دار پرداخت کنند (مرکز سرویس ۳) یا آنکه منتظر یکی از سه کارمند فروش شوند (مرکز سرویس ۲):



ویژگیهای سیستم‌های صف (ادامه): مکانیسم سرویس

- جزئیات بیشتری از مرکز سرویس ۲ در شکل مقابل نشان داده شده است.
- سه کارمند فروش به‌طور موازی به مشتریانی که در یک صف منتظرند سرویس می‌دهند.
- ممکن است که سرویس‌دهنده‌های موازی هر کدام به یک مشتری مجزا سرویس دهند که به این نوع سرویس‌دهی، **سرویس دسته‌ای** (batch service) گفته می‌شود.
- همچنین ممکن است که همه سرویس‌دهنده‌ها برای سرویس‌دهی به یک مشتری لازم باشند.

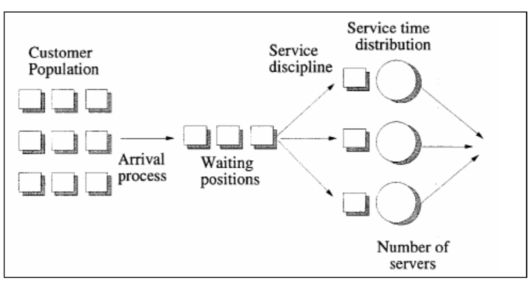


ویژگیهای سیستم‌های صف (ادامه)

- **انتظام سرویس (service discipline)**: یعنی رفتار سرویس‌دهنده با مشتریان منتظر در صف چگونه است و به چه ترتیبی یک مشتری را از بین آنها برای سرویس‌دهی انتخاب می‌شود.
  - منظور از انتظام سرویس آن است که سرویس دهنده از چه **الگوریتم زمانبندی** (scheduling algorithm) استفاده می‌کند؟
  - **مهمترین انواع انتظام سرویس عبارتند از:**
    - اولین ورود اولین خروج: First Come First Served (FCFS) یا First In First Out (FIFO)
    - به‌ترتیب تصادفی: Service In Random Order (SIRO)
    - کوتاه‌ترین زمان سرویس اول: Shortest Processing Time First (SPT)
    - بر اساس اولویت: Priority
    - آخرین زمان ورود اول: Last Come First Served (LCFS) یا List In First Out (LIFO)
    - با گردش نوبت: Round Robin (دارای اندازه کوانتم متناهی (finite quantum size))
    - اشتراک پردازنده: Processor Sharing (PS) (RR دارای اندازه کوانتم بی‌نهایت کوچک (infinitesimal))
    - کوتاه‌ترین زمان باقیمانده اول: Shortest Remaining Time First (SRTF)
    - سرویس‌دهنده‌های بی‌نهایت: Infinite Server (IS)

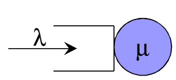
نمادهای مدلهای صف

- **یک سیستم صف مطابق شکل زیر را در نظر بگیرید:**



ویژگیهای سیستم صف
- فرآیند ورود
- توزیع زمان سرویس
- تعداد سرویس‌دهنده‌ها
- ظرفیت سیستم
- جمعیت سیستم
- انتظام سرویس

  - **به‌اختصار یک سیستم صف مجزا را به‌صورت زیر نشان می‌دهیم:**



نمادهای مدل‌های صف (ادامه)

- نماد (notation) استانداردی برای سیستم‌های صف وجود دارد که به نماد کِندال (Kendall's notation) معروف است.
- در این روش یک سیستم صف با شش مشخصه تعریف می‌شود: **A/S/m/B/K/SD**
- مشخصه‌های شش‌گانه مورد استفاده در نماد کندال عبارتند از:
  - **A**: فرایند ورود (یا توزیع زمانهای بین ورود)
  - **S**: توزیع زمان سرویس
  - **m**: تعداد سرویس‌دهنده‌ها
  - **B**: ظرفیت سیستم
  - **K**: اندازه جمعیت
  - **SD**: انتظام سرویس

نمادهای مدل‌های صف (ادامه)

- برای زمانهای بین ورود و سرویس مشخصه‌های زیر استفاده می‌شود:
  - M: توزیع نمایی (مارکوفی)
  - E_k: توزیع ارلنگ با پارامتر K
  - H_k: توزیع فوق نمایی (hyperexponential) با پارامتر K
  - D: قطعی (deterministic)
  - G: توزیع عمومی که با میانگین و واریانس مشخص می‌شود.
- دسته‌ای بودن سرویس یا ورود با بالانویس (superscript) مشخص می‌شوند:
  - مثلاً $M^{[x]}$ مشخص می‌کند که ورودیها نمایی بوده و هر بار گروهی به اندازه x وارد می‌شوند.
  - x طبق یک متغیر تصادفی مجزای دیگری مشخص می‌شود.
- اگر یکی از حروف مشخصه حذف شود به معنی آن است که مقدار پیش‌فرض استفاده می‌شود:
  - ظرفیت سیستم نامحدود است،
  - اندازه جمعیت نامتناهی است، یا
  - انتظام سرویس FCFS است.

نمادهای مدل‌های صف (ادامه)

- مثالهایی از نمادهای کندال:
  - مثال ۱: **M/D/5/40/200/FCFS**
    - زمانهای بین ورود طبق توزیع نمایی است.
    - زمانهای سرویس قطعی است.
    - پنج سرویس‌دهنده وجود دارند.
    - ظرفیت سیستم ۴۰ است که ۳۵ ظرفیت برای مکانهای انتظار وجود دارد.
    - کل جمعیت ۲۰۰ نفر هستند.
    - انتظام سرویس هم FCFS است.
  - مثال ۲: **M/M/1**
    - زمانهای بین ورود طبق توزیع نمایی است.
    - زمانهای سرویس طبق توزیع نمایی است.
    - یک سرویس‌دهنده وجود دارد.
    - ظرفیت سیستم نامحدود، جمعیت سیستم نامتناهی و انتظام سرویس FCFS است.

معیارهای کارایی سیستم های صف

- P_n (یا π_n): احتمال حالت پایدار وجود n مشتری در سیستم
- $P_n(t)$: احتمال وجود n مشتری در سیستم در زمان t
- λ: نرخ ورود
- λ_e: نرخ ورود موثر:
  - در مورد سیستم‌های دارای ظرفیت محدود مطرح است و مشخص کننده بخشی از فرآیند ورود است که فرصت ورود به سیستم را پیدا می‌کنند.
- μ: نرخ سرویس یک سرویس‌دهنده
- ρ: بهره‌وری سرویس‌دهنده
- A_n: زمان بین ورود مشتری (n-1)–ام و n–ام
- S_n: زمان سرویس n–امین مشتری وارد شده به سیستم

معیارهای کارایی سیستم‌های صف (ادامه)

- W_n: کل زمان صرف‌شده در سیستم توسط n–امین مشتری وارد شده به سیستم
- W_n^Q: کل زمان انتظار n–امین مشتری وارد شده به سیستم
- $L(t)$: تعداد مشتریان موجود در سیستم در زمان t
- $L_Q(t)$: تعداد مشتریان منتظر در صف در زمان t
- L: میانگین زمانی (time-average) تعداد مشتریان در سیستم در بلند مدت (long run)
- L_Q: میانگین زمانی تعداد مشتریان منتظر در صف در بلند مدت
- w: میانگین زمان صرف‌شده برای هر مشتری در بلند مدت
- w_Q: میانگین زمان انتظار هر مشتری در صف در بلند مدت

میانگین زمانی تعداد مشتریان در سیستم

- **یک سیستم صف را در طی یک بازه زمانی T در نظر بگیرید:**
  - فرض کنید که T_i نشان دهنده بخشی از زمان در بازه $[0,T]$ که دقیقاً i مشتری در سیستم وجود داشته‌اند. در این صورت میانگین زمانی وزن‌دار تعداد (time-weighted-average number) مشتریان در سیستم به‌صورت زیر تعریف می‌شود:

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i\left(\frac{T_i}{T}\right)$$

  - اگر مساحت سطح زیر منحنی $L(t)$ را در نظر بگیریم خواهیم داشت:

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_0^T L(t)dt$$

  - در این صورت میانگین زمانی بلند مدت تعداد مشتریان در سیستم به‌صورت زیر تعریف می‌شود:

$$\hat{L} = \frac{1}{T}\int_0^T L(t)dt \to L \quad \text{as} \quad T \to \infty$$

  - پارامترهای هت‌دار (مثل $\hat{L}$) حاصل داده‌های مشاهده و اندازه‌گیری هستند.

قانون لیتل در مورد صف مجزا

- قانون لیتل در مورد صف مجزا برقرار است:

میانگین تعداد مشتریان

$$\hat{L} = \hat{\lambda}\hat{w}$$

میانگین زمان صرف‌شده در سیستم

نرخ ورود

$$L = \lambda w \quad \text{as} \quad T \to \infty \ \text{ and } \ N \to \infty$$

- این قانون در مورد همه انواع صفها (بدون توجه به تعداد سرویس‌دهنده‌ها، انتظام سرویس یا سایر مشخصات) صدق می‌کند.

قانون لیتل در مورد صف مجزا (ادامه)

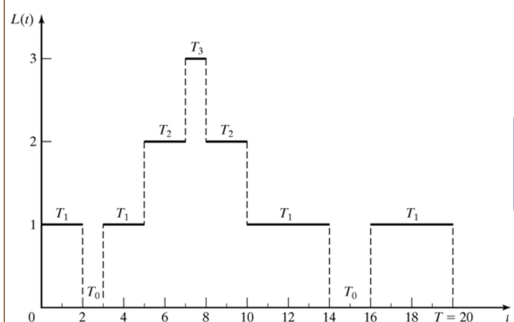
- مثالی از کاربرد قانون لیتل در مورد یک صف G/G/1/N/K:
  - به‌طور میانگین در هر ۴ واحد زمان یک ورود به چنین صفی انجام می‌شود و هر مشتری وارد شده به‌طور میانگین 4.6 واحد زمان را در سیستم صرف می‌کند. بنا بر این تعداد مشتریان موجود در سیستم در هر زمان به‌صورت زیر محاسبه می‌شود:
    - $E(A) = 1/\lambda = 4 \quad => \lambda = 1/4$
    - $W = 4.6$
    - $L = \lambda W = (1/4)(4.6) = 1.15$ مشتری

میانگین زمانی تعداد مشتریان در صف (ادامه)

- میانگین زمانی وزن‌دار تعداد مشتریان منتظر در صف به‌صورت زیر بدست می‌آید:

$$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)dt \to L_Q \quad \text{as} \quad T \to \infty$$

- مثال: یک صف G/G/1/N/K را (که N>3 و K>4) در نظر بگیرید:



$$\hat{L} = [0(3)+1(12)+2(4)+3(1)]/20$$
$$= 23/20 = 1.15 \text{ customers}$$

$$L_Q(t) = \begin{cases} 0, & \text{if } L(t)=0 \\ L(t)-1, & \text{if } L(t)\geq 1 \end{cases}$$

$$\hat{L}_Q = \frac{0(15)+1(4)+2(1)}{20} = 0.3 \text{ customers}$$

میانگین زمان صرف‌شده در سیستم

- **میانگین زمان صرف‌شده در سیستم** (average time spent in system) برای هر مشتری که به آن **زمان سیستمی میانگین** (average system time) به‌صورت زیر محاسبه می‌شود:

$$\hat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i$$

که در آن $W_1, W_2, \ldots, W_N$ زمانهای مجزایی است که هر کدام از N مشتری در طی بازه [0,T] در سیستم صرف نموده‌اند.

- برای سیستم‌های پایدار داریم:

$$\hat{w} \to w \text{ as } N \to \infty$$

- اگر تنها بخش صف سیستم را در نظر بگیریم خواهیم داشت:

$$\hat{w}_Q = \frac{1}{N}\sum_{i=1}^{N} W_i^Q \to w_Q \quad \text{as} \quad N \to \infty$$

- سیستم صف G/G/1/N/K را دوباره در نظر بگیرید. زمان سیستمی میانگین به‌صورت زیر محاسبه می‌شود:

$$\hat{w} = \frac{W_1 + W_2 + \ldots + W_5}{5} = \frac{2 + (8-3) + \ldots + (20-16)}{5} = 4.6 \text{ time units}$$

بهره‌وری سرویس‌دهنده

- همانطوری که قبلاً تعریف کردیم، بهره‌وری به بخشی از زمان گفته می‌شود که سرویس‌دهنده مشغول است:
  - در این صورت بهره‌وری مشاهده شده که با $\hat{\rho}$ نشان داده می‌شود در طی یک بازه زمانی $[0,T]$ تعریف می‌شود.
  - در این صورت بهره‌وری سرویس‌دهنده در بلندمدت ρ خواهد بود.
  - برای سیستم‌هایی که به حالت پایدار رسیده‌اند خواهیم داشت:

$$\boxed{\hat{\rho} \to \rho \quad \text{as} \quad T \to \infty}$$

بهره‌وری سرویس‌دهنده

- صف G/G/1 را در نظر بگیرید:
  - **چنین سیستم صفی دارای میانگین نرخ ورود λ مشتری در واحد زمان، میانگین زمان سرویس $E(S) = 1/\mu$ واحد زمان، ظرفیت سیستم نامحدود و جمعیت مشتریان نامتناهی است.**
  - **قانون لیتل در مورد چنین سیستمی قابل استفاده است: $L = \lambda w$**
  - **برای یک سیستم پایدار، میانگین نرخ ورود به سرویس‌دهنده، λ_s، باید مساوی λ باشد.**
    - چون سیستم به حالت پایدار رسیده: $\lambda=X$. از طرفی، خروجی سیستم و خروجی سرویس‌دهنده با هم مساوی هستند $(X_s=X)$. همچنین، چون سرویس‌دهنده به حالت پایدار رسیده، بخش سرویس‌دهنده هم باید به حالت پایدار رسیده باشد. بنا بر این باید $\lambda_s=X_s$. در نتیجه می‌توان نتیجه گرفت که: $\lambda_s=X_s=X=\lambda$.
  - **میانگین تعداد مشتریان در سرویس‌دهنده:**

$$\boxed{\hat{L}_s = \frac{1}{T}\int_0^T \big(L(t) - L_Q(t)\big)dt = \frac{T - T_0}{T}}$$

  - مطابق مطالب جلسه ۴ (شبیه‌سازی):

$$\boxed{\hat{L}_s = \frac{1}{T}\int_0^T B(t)dt = \frac{B}{T}}$$

بهره‌وری سرویس‌دهنده

- در حالت کلی برای یک صف تک‌سرویس‌دهنده‌ای (با توجه به تعریف بهره‌وری):

$$\hat{L}_s = \hat{\rho} \rightarrow L_s = \rho \quad \text{as } T \rightarrow \infty$$
$$\Rightarrow \quad \rho = \lambda E(s) = \frac{\lambda}{\mu}$$

- قانون لیتل هم به کل سیستم صف $(L = \lambda W)$، هم به بخش صف انتظار $(L_Q = \lambda W_Q)$ و هم به بخش سرویس‌دهنده $(\rho = \lambda E(S))$ قابل اعمال است.

- برای اینکه یک صف تک‌سرویس‌دهنده‌ای در حالت پایدار باشد باید:

$$\rho = \frac{\lambda}{\mu} < 1$$

- اگر $\lambda > \mu$ باشد صف ناپایدار بوده و به معنی آن است که بهره‌وری سرویس‌دهنده **یک** بوده و همیشه **مشغول** است.

بهره‌وری سرویس‌دهنده

- صف G/G/m را در نظر بگیرید:
  - چنین سیستم صفی دارای m سرویس‌دهنده یکسان است که به‌طور موازی کار می‌کنند.
  - اگر یک مشتری وارده بیش از یکی از سرویس‌دهنده‌ها را بیکار ببیند، مشتری یکی از سرویس‌دهنده‌ها را بدون قایل شدن هرگونه تفاوتی مابین سرویس‌دهنده‌ها انتخاب می‌کند.
  - برای سیستم‌هایی که در حالت تعادل آماری (statistical equilibrium) (یا حالت پایدار) هستند، میانگین تعداد سرویس‌دهنده‌های مشغول، L_s، عبارت است از:

$$L_s = \lambda E(s) = \lambda/\mu$$

  - میانگین بهره‌وری سرویس‌دهنده در بلند مدت عبارت است از:

$$\rho = \frac{L_s}{m} = \frac{\lambda}{m\mu}$$

که برای آن شرط پایداری سیستم $\lambda < m\mu$ است.

بهره‌وری سرویس‌دهنده و کارایی سیستم

- کارایی سیستم با توجه به یک بهره‌وری داده شده ρ تغییر می‌کند.
  - برای مثال، برای یک صف D/D/1 را که $E(A) = 1/\lambda$ و $E(S) = 1/\mu$ است داریم:

$$L = \rho = \lambda/\mu, \quad w = E(S) = 1/\mu, \quad L_Q = W_Q = 0$$

  - در این حالت هیچ گونه صف انتظاری تشکیل نمی‌شود و با تغییر λ و μ، بهره‌وری سرویس‌دهنده می‌تواند هر مقداری بین صفر و یک باشد.
- در حالت کلی، تغییر زمانهای بین ورود و سرویس باعث می‌شود که طول صف انتظار کم و زیاد شود.

بهره‌وری سرویس‌دهنده و کارایی سیستم

- مثال: یک پزشک که بیماران را برای هر ده دقیقه زمانبندی می‌کند، زمان S_i را برای i-امین مشتری صرف می‌کند:

$$S_i = \begin{cases} 9 \text{ min}, p = 0.9 \\ 12 \text{ min}, p = 0.1 \end{cases}$$

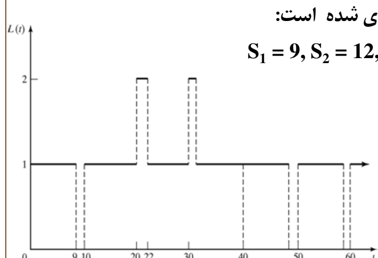
  - ورودها قطعی هستند: $A_1 = A_2 = \ldots = \lambda^{-1} = 10$
  - زمانهای سرویس تصادفی با میانگین $E(S) = 9.3$ min و واریانس $V(S) = 0.81$ min² هستند.
  - بهره‌وری پزشک به‌طور میانگین برابر خواهد بود با:

$$\rho = \lambda/\mu = 0.93 < 1.$$

  - در نظر بگیرید که سیستم با زمانهای سرویس زیر شبیه‌سازی شده است:

$$S_1 = 9, S_2 = 12, S_3 = 9, S_4 = 9, S_5 = 9, \ldots.$$

  - در این صورت نتایج نمودار مقابل بدست آمده است:
    - مشاهده می‌شود که یک زمان سرویس به نسبت طولانی $(S_2 = 12)$ منجر به تشکیل شدن موقتی یک صف انتظار می‌شود.



رفتار حالت پایدار مدل‌های مارکوفی دارای جمعیت نامتناهی

- مدل‌های مارکوفی (Markovian models)، مدل‌هایی هستند که در آنها:
  - فرآیند ورود طبق توزیع نمایی با یک نرخ ورود مثل λ است.
  - اما زمان سرویس ممکن است که دارای توزیع نمایی (M) یا هر توزیع دلخواه (G) باشد.
- یک سیستم صف دارای تعادل آماری (statistical equilibrium) است اگر احتمال اینکه در یک حالت داده شده است وابسته به زمان نباشد. یعنی در حالت تعادل یا پایدار (steady state) احتمالات متناظر با حالتهای مختلف ثابت است. این احتمالات به‌صورت زیر تعریف می‌شوند:

$$P[L(t)=n\ ] = P_n(t) = P_n$$

- راه حلهای ریاضی برای حل تقریبی حالت پایدار مدل‌های صف وجود دارند.
- همچنین، شبیه‌سازی حالت پایدار نیز برای تحلیل سیستم‌های صف پیچیده و بزرگ که راه حلی ریاضی ندارند استفاده می‌شود.

رفتار حالت پایدار مدل‌های مارکوفی دارای جمعیت نامتناهی (ادامه)

- برای مدل‌های صف مجزا، میانگین تعداد مشتریان در سیستم در حالت پایدار (L) به صورت زیر محاسبه می‌شود:

$$\boxed{L=\sum_{n=0}^{\infty} nP_n}$$

- با اعمال قانون لیتل به کل سیستم و نیز به بخش صف خواهیم داشت:

$$\boxed{\begin{array}{c} w=\dfrac{L}{\lambda},\quad w_Q = w-\dfrac{1}{\mu} \\ L_Q=\lambda w_Q \end{array}}$$

رفتار حالت پایدار صفهای M/G/1

- صف M/G/1 یک صف تک‌سرویس‌دهنده‌ای دارای ورودهای پواسان و ظرفیت نامحدود است.
- فرض کنید که زمانهای سرویس دارای میانگین $1/\mu$ و واریانس σ^2 بوده و همچنین $\rho = \lambda/\mu < 1$. در این صورت پارامترهای حالت پایدار صف M/G/1 به‌صورت زیر خواهند بود:

$$\rho = \lambda/\mu, \quad P_0 = 1-\rho$$
$$L = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}, \quad L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$
$$w = \frac{1}{\mu} + \frac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}, \quad w_Q = \frac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$$

رفتار حالت پایدار صفهای M/G/1 (ادامه)

- عبارت ساده‌ای برای احتمالات حالت پایدار $P_0, P_1, \ldots$ صف M/G/1 وجود ندارد.
- میانگین تعداد مشتریان در حال سرویس $L - L_Q = \rho$ که همان بهره‌وری سرویس‌دهنده است.
- میانگین طول صف، L_Q، می‌تواند به‌صورت زیر نوشته شود:

$$L_Q = \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2\sigma^2}{2(1-\rho)}$$

  - اگر λ و μ ثابت نگهداشته شوند میانگین طول صف، L_Q، به تغییر یا واریانس، σ^2، زمانهای سرویس بستگی خواهد داشت.

رفتار حالت پایدار صفهای M/G/1 (ادامه)

- **مثال:** دو کارگر به نامهای Able و Baker برای یک کار با هم رقابت می‌کنند. Able ادعا می‌کند که به‌طور میانگین از Baker سریعتر است. اما Baker ادعا می‌کند که روش کارش سازگارتر بوده و خیلی تغییرات (کم و زیاد شدن) در زمان سرویس ندارد.
  - ورودیها پواسان با نرخ $\lambda = 2$ در ساعت (یا 1/30 در دقیقه) هستند.
  - برای Able داریم: $1/\mu = 24$ دقیقه و $\sigma^2 = 20^2 = 400\ min^2$ لذا خواهیم داشت:

$$L_Q = \frac{(1/30)^2[24^2 + 400]}{2(1-4/5)} = 2.711\ \text{customers}$$

  - بخشی از زمان که ورودیها Able را بیکار می‌بینند و لذا تاخیری را متحمل نمی‌شوند:

$$P_0 = 1-\rho = 1/5 = 20\%$$

  - برای Baker داریم: $1/\mu = 25$ دقیقه و $\sigma^2 = 2^2 = 4\ min^2$. لذا خواهیم داشت:

$$L_Q = \frac{(1/30)^2[25^2 + 4]}{2(1-5/6)} = 2.097\ \text{customers}$$

  - بخشی از زمان که ورودیها Baker را بیکار می‌بینند و لذا تاخیری را متحمل نمی‌شوند:

$$P_0 = 1-\rho = 1/6 = 16.7\%$$

  - بنا بر این با وجودی که به‌طور میانگین Able سریعتر است، چون میزان تغییرات زمان سرویسش بیشتر از Baker است به طور میانگین طول صف آن ۳۰٪ بیشتر است.

رفتار حالت پایدار صفهای M/M/1

- **اگر زمانهای سرویس در صف M/G/1 دارای توزیع نمایی با میانگین $1/\mu$ باشند، آنگاه واریانس $\sigma^2 = 1/\mu^2$ خواهد بود. چنین صفی M/M/1 است.**
  - **صف M/M/1 یک مدل تقریبی سودمندی است که زمانهای سرویس دارای انحراف معیاری تقریباً برابر با میانگین‌شان هستند.**
  - **پارامترهای حالت پایدار صف M/M/1 عبارتند از:**

$$\rho = \lambda/\mu, \quad P_n = (1-\rho)\rho^n$$

$$L = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}, \quad L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}$$

$$w = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}, \quad w_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}$$

رفتار حالت پایدار صفهای M/M/1 (ادامه)

- مثال: یک صف M/M/1 دارای نرخ سرویس μ=10 مشتری در ساعت است.
  - نحوه تغییرات L و w را با افزایش نرخ ورود، λ، از 5 به 8.64 به افزایشهای ۲۰٪ در نظر بگیرید:

λ	5.0	6.0	7.2	8.64	10.0
ρ	0.500	0.600	0.720	0.864	1.000
L	1.00	1.50	2.57	6.35	∞
w	0.20	0.25	0.36	0.73	∞

  - اگر $\lambda/\mu \geq 1$ باشد طول صف انتظار به‌طور پیوسته افزایش می‌یابد.
  - افزایش زمانی سیستمی میانگین (w) و تعداد میانگین در سیستم (L) یک تابع غیرخطی از ρ است.

صفهای دارای چند سرویس‌دهنده

- نماد کندال چنین صفی M/M/m است که m کانال به‌طور موازی عمل می‌کنند.
  - هر کانال دارای توزیع زمان سرویس مستقل و یکسان (iid) نمایی با میانگین $1/\mu$ است.
  - برای حصول تعادل آماری، بار (load) پیشنهادی (λ/μ) باید در رابطه $\lambda/\mu < m$ صدق کند که بهره‌وری سرویس‌دهنده $\rho = \lambda/(m\mu)$ خواهد بود.
  - برخی از احتمالات حالت پایدار این سیستم عبارتند از:

$$\rho = \lambda / m\mu$$

$$P_0 = \left\{ \left[\sum_{n=0}^{m-1} \frac{(\lambda/\mu)^n}{n!} \right] + \left[\left(\frac{\lambda}{\mu} \right)^m \left(\frac{1}{m!} \right) \left(\frac{m\mu}{m\mu - \lambda} \right) \right] \right\}^{-1}$$

$$L = m\rho + \frac{(m\rho)^{m+1} P_0}{m(m!)(1-\rho)^2} = m\rho + \frac{\rho \, P\big(L(\infty) \geq m\big)}{1-\rho}$$

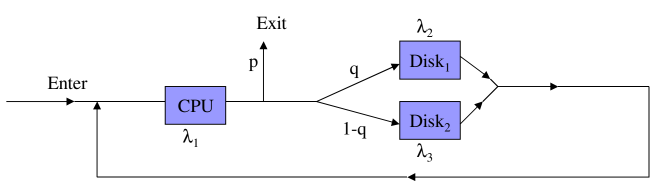
$$w = \frac{L}{\lambda}$$

صفهای دارای چند سرویس‌دهنده (ادامه)

- مدل‌های صف چندسرویس‌دهنده عمومی دیگر عبارتند از:
  - M/G/m:
    - توزیع زمانهای سرویس عمومی و تعداد سرویس‌دهنده‌های موازی m است.
    - پارامترهای این صف قابل تقریب از مدل M/M/m هستند.
  - M/G/∞:
    - توزیع زمانهای سرویس عمومی و تعداد سرویس‌دهنده‌های موازی نامتناهی است.
  - M/M/m/N:
    - زمانهای سرویس دارای توزیع نمایی با نرخ μ و تعداد سرویس‌دهنده‌ها m است.
    - ظرفیت کل سیستم $N \geq m$ مشتری است.

کاربردهای مدلهای صف در ارزیابی کارایی سیستم‌های کامپیوتری

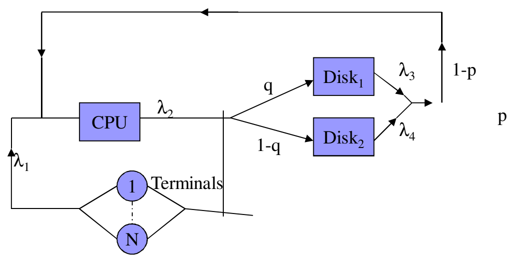
- مدل‌های صف دارای کاربرد وسیعی در مدل‌سازی و ارزیابی سیستم‌ها و از جمله سیستم‌های کامپیوتری هستند.
- برای مثال یک سیستم کامپیوتری را در نظر بگیرید:
  - کارها (jobs) برای اجرا وارد یک سیستم کامپیوتری می‌شوند. توسط یک پردازنده (CPU) با نرخ λ_1 پردازش می‌شوند، با احتمال p از سیستم خارج می‌شوند، یا آنکه ممکن است با یک احتمال q با یکی از دو دیسک و با احتمال 1-q با دیسک دیگر کار داشته باشند.
  - در هر کدام از زیرسیستم‌های CPU، $Disk_1$ و $Disk_2$، صف‌هایی برای کارها تشکیل می‌شود.



  - مدل فوق، مثالی از یک شبکه صف باز (open queueing network) است. چون از بیرون به آن ورودی داریم.

کاربردهای مدلهای صف در ارزیابی کارایی سیستم‌های کامپیوتری (ادامه)

- مثالی از یک شبکه صف بسته (closed queueing network):
  - یک سیستم اشتراک زمانی را که متشکل از تعدادی ترمینال و یک سیستم مرکزی متشکل از یک CPU و دو دیسک را در نظر بگیرید:



  - مدل فوق، مثالی از یک شبکه صف بسته است. چون تعداد کارهای موجود در سیستم به تعداد ترمینالها بوده و ثابت است.

کاربردهای مدلهای صف در ارزیابی کارایی سیستم‌های کامپیوتری (ادامه)

- با استفاده از مدل‌های صف و شبکه‌های صف می‌توانیم سیستم‌های فوق را تحلیل کنیم.
- یعنی مثلاً برای ارزیابی کارایی این سیستم‌ها از نتایج حل سیستم‌های صف استفاده کنیم.
- اگر سیستم یا شبکه صف مورد استفاده راه‌حل تحلیلی نداشته باشد، آنرا شبیه‌سازی می‌کنیم.
- در جلسات سوم و چهارم با شبیه‌سازی سیستم‌های صف آشنا شدیم.
- در جلسات بعدی ابتدا در مورد روشهای تحلیل مدل‌های صف مجزا و سپس شبکه‌های صف صحبت می‌کنیم.