EECS 440 System Design of a Search Engine

Winter 2021

Lecture 3: Project planning

Nicole Hamilton
https://web.eecs.umich.edu/~nham/
nham@umich.edu

Agenda

1. Course details.
2. Your project.
3. Project planning.

Agenda

1. Course details.
2. Your project.
3. Project planning.

details

1. Anyone still looking for a group?
2. Homework 2 posted. You will need to install GanttProject.

Agenda

1. Course details.
2. Your project.
3. Project planning.

Overall plan

1. Get you crawling ASAP.
2. Cover the OS topics you'll need.
3. Explain how each piece works just in time for you to build it, starting with the crawler.

Your project

I've broken your engines into 8 levels of functionality:

0 Basic plan for your project.
1 Parse text files into a hash table.
2 Build a crawler.
3 Build a reverse word index.
4 Create a user interface.
5 Build a constraint solver and query parser.
6 Build a ranker.
7 Advanced functionality.

Level 0: Create a plan

1. Decide what information you'll collect.
2. Decide whether to keep a copy of every document you crawl and index.
3. Estimate and acquire needed storage.
4. Break the problem down into pieces. Decide what it does and how it works and estimate LOC.
5. Decide how you'll work as a team, meetings, source control, common libraries, etc.
6. Prepare a written plan (which you will submit) with deliverables, milestones and responsibilities.

Level 1: Parse text files into a hash table

1. Parse text files into tokens, stripping out most HTML but identifying links and anchor text. (HW3)
2. Traverse a directory on disk, indexing the content in a hash table in memory.
3. Given a search work, list the occurrences by searching the index.

Level 2: Build a crawler

1. Given a list of URLs, retrieve the documents by HTTP and HTTPS.
2. Create a crawler to manage a frontier of URLs to be explored, eliminating URLs already seen.
3. Ensure it can be restarted.
4. Decide what makes one URL better than another. Be polite.
5. Define a seed list to start crawling.
6. Parallelize the tasks of crawling, parsing and index build.

Level 3: Build reverse word index

1. Define a file format. Decide how you will merge multiple index files.
2. Decide how you will structure and encode your data.
3. Decide how to number locations across documents. Accumulate useful statistics.
4. Demonstrate you can build the index.
5. Demonstrate you read the index.
6. Create index stream readers for words and documents.

Level 4: Create a user interface

1. Initially, a simple command line interface.
2. A simple HTTP server to act as a wrapper UI.
3. Report title, clickable URL, possibly other interesting information.

Level 5: Constraint solver and query parser

1. Create derive AND stream reader.
2. Demonstrate ability to find next document containing set of words.
3. Derive additional stream readers for OR, phrase and NOT.
4. Build a top-down recursive descent parser to compile queries into index stream readers.
5. Possibly support stemming and stop words.
6. Demonstrate searching on complex queries.

Level 6: Build a ranker

1. Rank using a simple bag-of-words technique.
2. Rank using static attributes of the page.
3. Rank using heuristics that consider the quality of match, e.g., exact phrases, words found in the title or URL, etc.
4. Demonstrate ability to produce a useful “10 best” list of search results.

Level 7: Advanced functionality

1. Create a training set and use gradient descent to improve your relevance.
2. Rank using a neural net or other ML technique.
3. Create a snippet to go with a reported hit.
4. Associate anchor text with the document it describes.
5. Add PageRank.
6. Distribute crawling, indexing or query processing across a small network of machines.
7. Index PDFs.

Project plan

Due February 14, 2021

1. Your project plan must satisfy functionality level 0.
2. Presentation and demo by April 19
3. Final report by April 21.
4. Final review with me by April 26.

Project plan

Due February 14, 2021

5. Your title page must include your group photo.
6. Identify your group and tell me about yourselves.
7. Briefly outline how your group formed and how you expect to work together as a group.

Project plan

Due February 14, 2021

8. Outline your basic technical strategy.
9. Identify what information you will collect, how much storage you'll need, and where you'll get it.
10. Break your engine down into major pieces, list what each will do, and identify the inputs and outputs.

Project plan

Due February 14, 2021

11. Identify which pieces will run as separate processes or threads, which will need to be restarted after a crash, etc.
12. Create a high-level diagram of your engine.
13. Estimate the lines of code you think you may needed for each piece and who will do it.

Project plan

Due February 14, 2021

14. Create a Gantt chart of the activities.
15. Identify a major milestone by March 15.
16. Identify what you expect to have at the milestone and in your final product.

Organizing your team

1. Success on this project is critically tied to your ability to work cooperatively and effectively as a team.
2. You will need to come to decisions quickly and they will need to be based on consensus, not who's in charge.
3. You will need to involve everyone all the time. You will all need to step up to do right by your teams.
4. Decide how you'll organize the work, e.g., should everyone code their own tests or should you test each others' work?
5. Check in often.
6. Questions? Pitfalls?

Agenda

1. Course details.
2. Search basics: Ranking
3. Your project.
4. Project planning.

Problem-solving versus problem-choosing

A fabulous solution to a problem nobody cares about is still a solution nobody cares about.

But even a mediocre solution to a really important problem can be important.

A lot of *problem-solving* is coming up with ideas of how to do something.

A lot of *problem-choosing* is deciding *what you won't do.*

Planning

Planning in industry is usually done in two parts:

1. A business plan or market requirements document (MRD) that defines the objectives and why they're worthwhile.
2. A project plan for accomplishing the objectives.

Usually starts with an opportunity.

Typical business plan
Table of contents

Executive summary

1. Opportunity and market analysis
2. The solution and concept
3. Marketing and sales
4. Product development and operations
5. Team and organization
6. Risks
7. Financial plan

Appendix: Detailed financial plan

Executive summary

1. Why is this a problem and why are customers willing to pay?
2. How will the problem be solved?
3. Why is the venture uniquely positioned to do this?
4. What are the economics? Is this a growth opportunity?
5. Who's on the team and what partnerships are already in place?

Opportunity and market

1. What is the problem or need being solved?
2. Who is the customer?
3. How large is the total addressable market and how is it growing?
4. Is the current market context favorable?

Solution and concept

1. What is the product or service?
2. What does a day-in-the-life look like for a customer before and after?
3. Which customers have validated the product and are willing to pay for it?
4. What is unique and defendable?
5. What is the business and economic model? What are the margins?

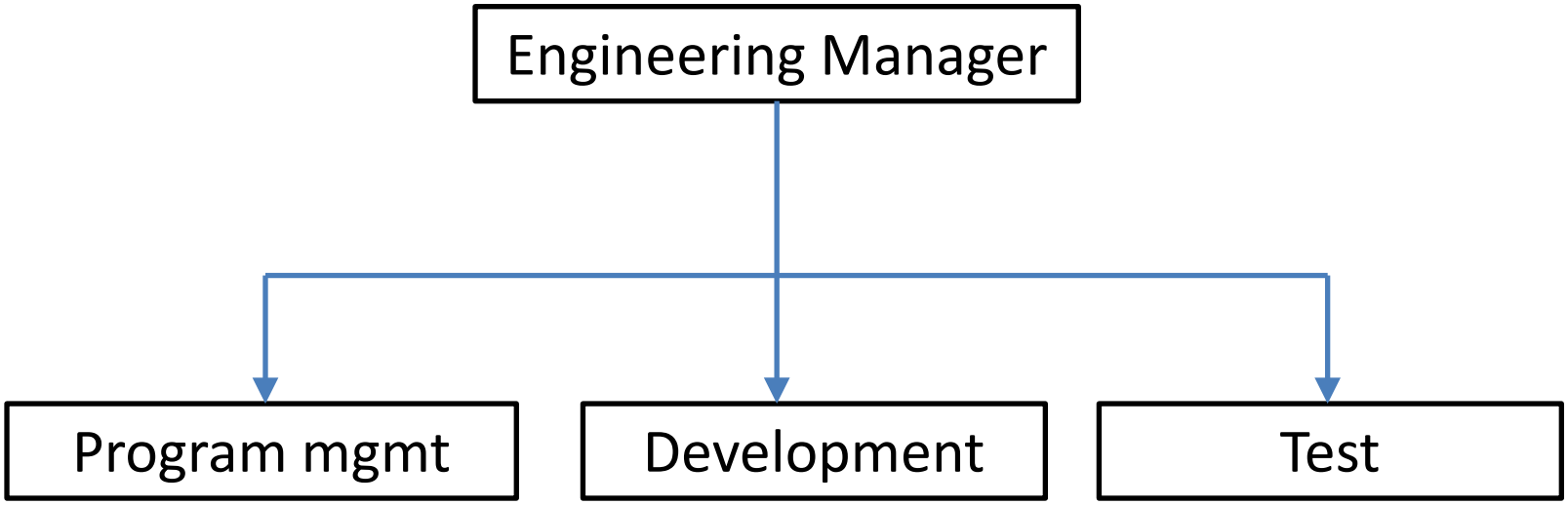
Marketing and sales

1. What are the appropriate marketing mediums to reach customers?
2. What is the most appropriate sales channel?
3. Who are the decision makers and who are the influencers?
4. How long is the expected sales cycle?
5. Are there opportunities for partnerships to advertise and sell?

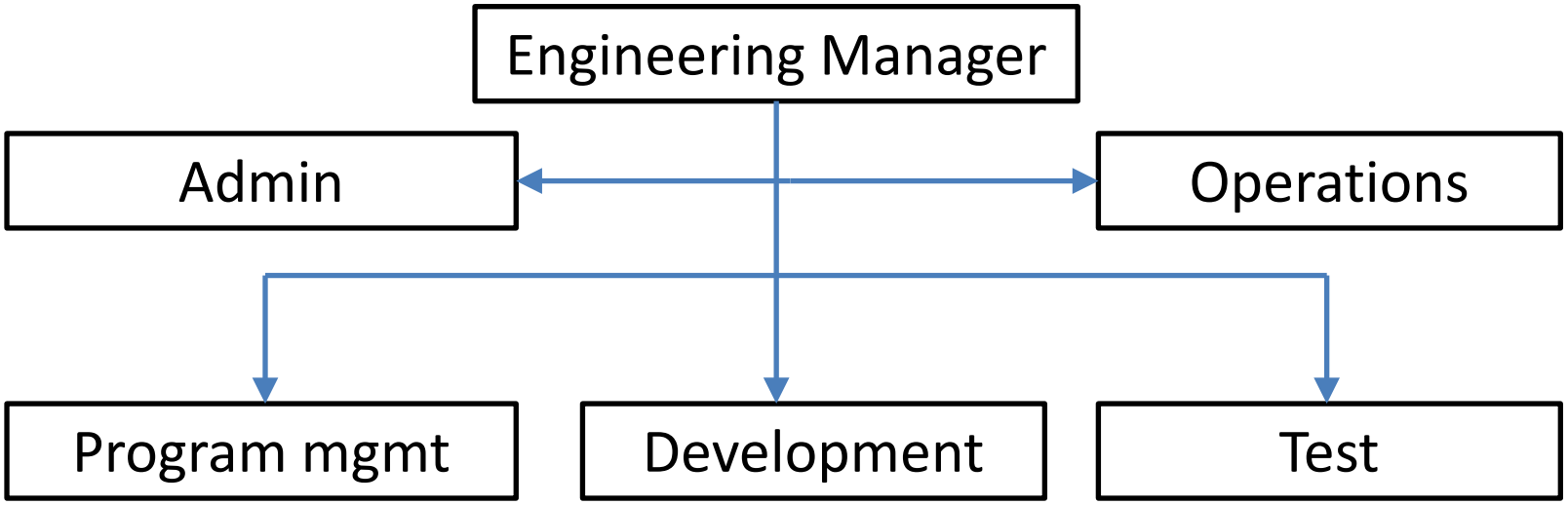
Product development

1. What is the current state of any product development?
2. What will be required to complete and ship products?
3. What are timelines and milestones?
4. What are the key risks?
5. What is the value chain for production and product delivery?
6. Are there any patent, trade secret or other advantages?
7. Are there any regulatory hurdles?

Typical dev organization



Typical dev organization



Team

1. What are the backgrounds and roles of the founders and early employees?
2. What are the team's passions and skills and why are they committed?
3. What key hires are needed?
4. What head count is needed, by function?
5. Who are the advisors and board members?

Risks

1. What are the key product development risks and external dependencies?
2. What is being done to mitigate risk?
3. Who are your main competitors and how are you differentiated from them?
4. Can large players easily enter the market? Are there any substitutes?
5. What strategies can be used to mitigate competitive threats?

Appendix: Detailed financial plan

1. Five-year detailed cash flow, income and balance sheets.
2. Financial assumptions, e.g., customer penetration rates, pricing.
3. Are purchasing decisions cyclical?
4. What are the largest costs, e.g., engineering, regulatory trials, manufacturing or marketing?
5. How will the product and sales costs change as volume grows?
6. Has customer support and maintenance been factored in?

Typical project plan

1. Introduction and description of the project.
2. Goals and objectives.
3. Scope.
    1. Scope definition.
    2. Items beyond the scope.
    3. Risk assessment.
4. Project strategy.
    1. Architecture and block diagrams.
    2. Sizing estimates, e.g., LOC.
    3. Roles and responsibilities.
    4. Timeline.
    5. Deliverables.
    6. Milestones.

Project planning

1. Decide your objectives.
2. Decompose it into pieces.
3. Estimate the size of each piece and how you'll do it.
4. Layout a timeline for the project.
5. Specify deliverables and milestones with *exit criteria*.
6. Measure your progress against the plan.

Decomposition

To plan a complex project, you break it down into understandable pieces.

You may not know how to do each piece but you can try to list them.

Decomposition

To plan a complex project, you break it down into understandable pieces.

You may not know how to do each piece but you can try to list them.

You keep doing this recursively.

When a piece is small enough, you estimate the LOC it will require.

Add up the pieces, use that to decide how long it will take.

Gantt charts

Project timelines usually laid out a Gantt chart, named after the inventor, Henry Gantt (1861 to 1919).

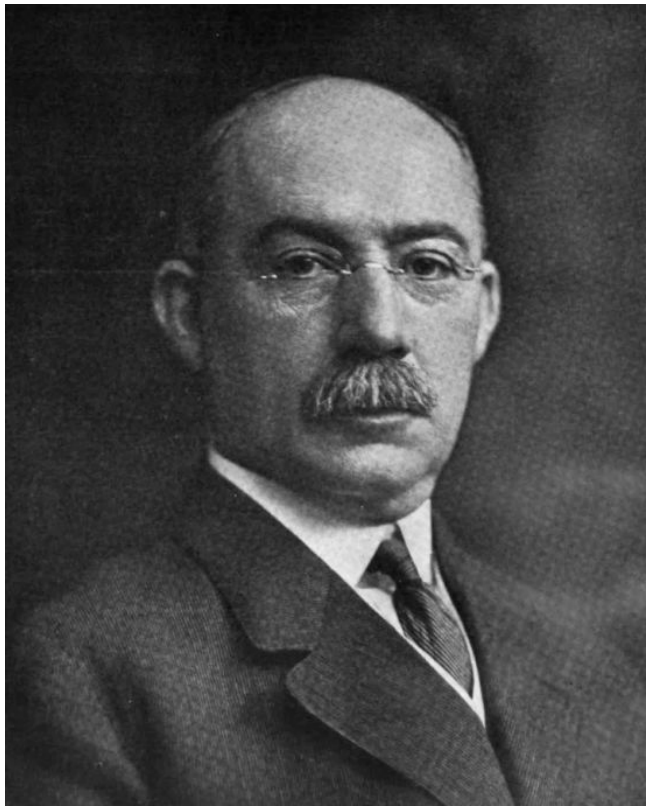
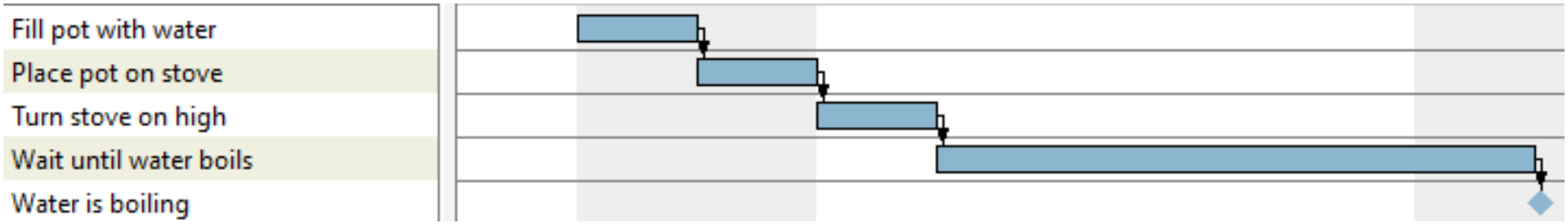


Image source: https://en.wikipedia.org/wiki/Henry_Gantt#/media/File:Henry_L._Gantt.jpg

Gantt charts

Bar chart showing a project schedule.
Tasks listed vertically, horizontal axis is time.
Bars represent the duration of each task.

Example project: Boiling water

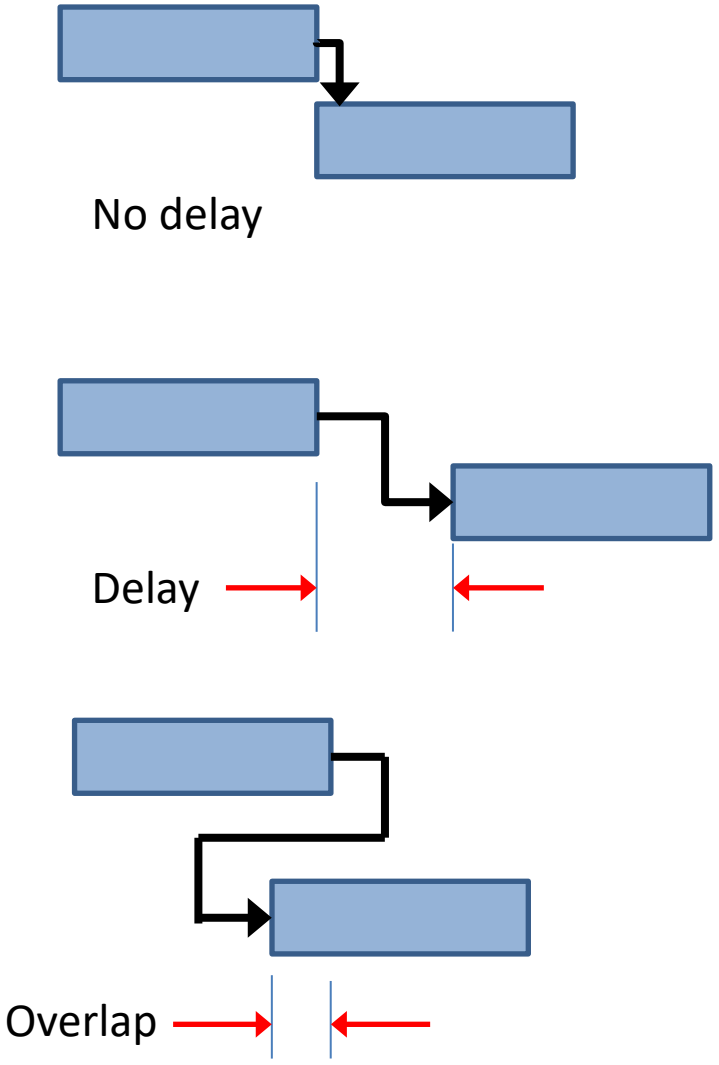


Most popular project planning software is Microsoft Project but it's not free.
A good free alternative is GanttProject from https://www.ganttproject.biz/.

Dependencies

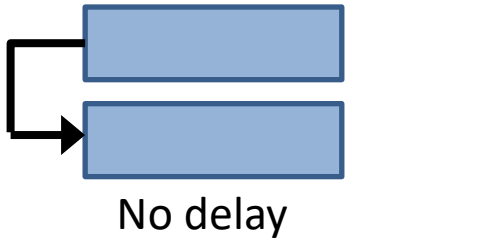
Dependencies between tasks represented by arrows between them.

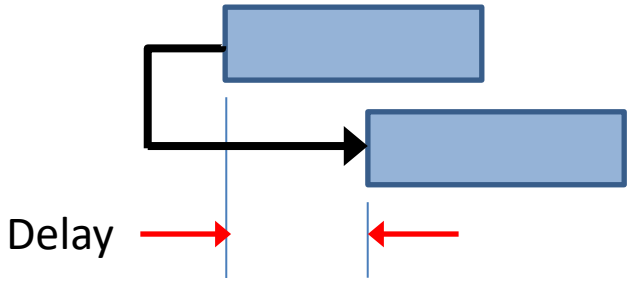
Shown here: Finish-to-start dependencies.

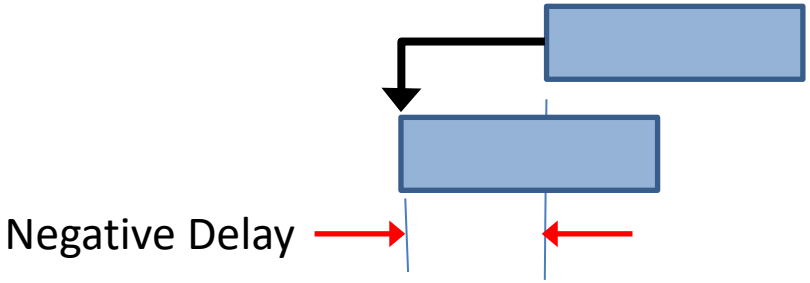


Dependencies

Start-to-start dependencies.

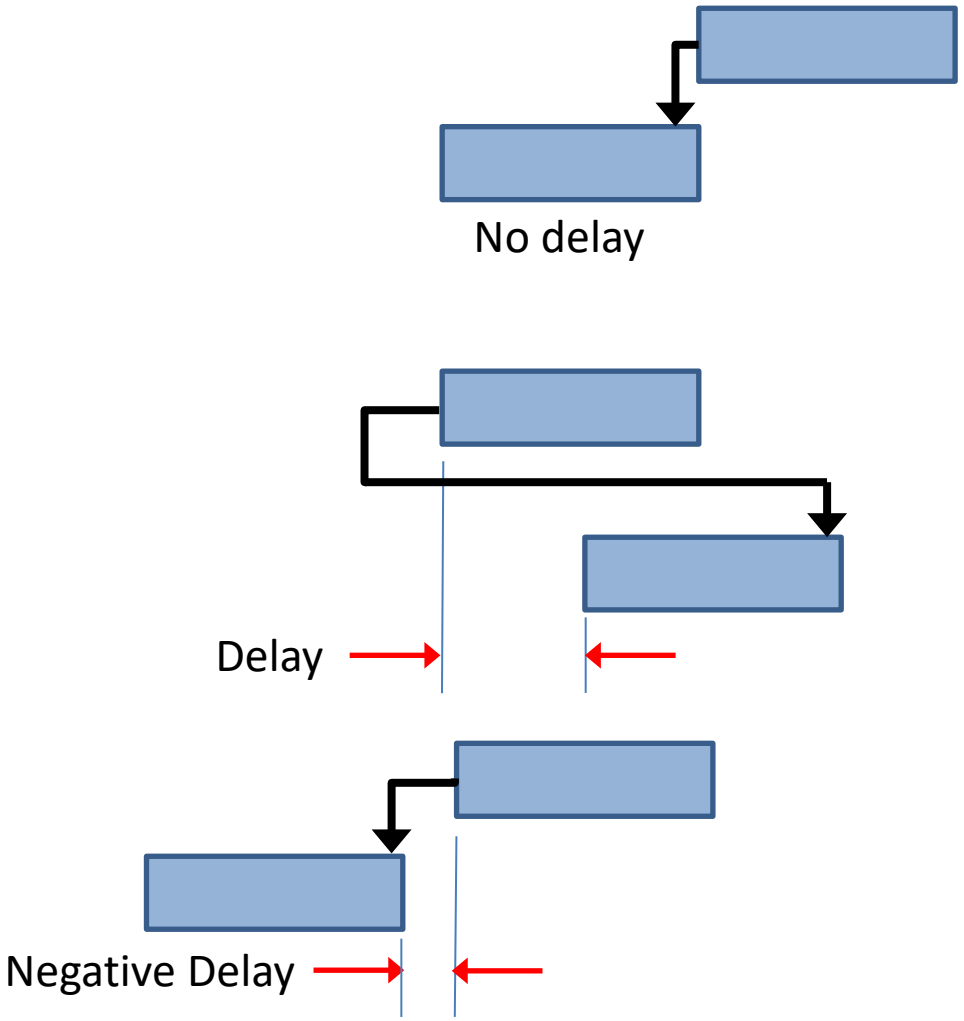






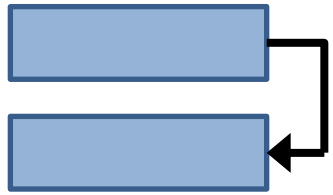
Dependencies

Start-to-finish dependencies.

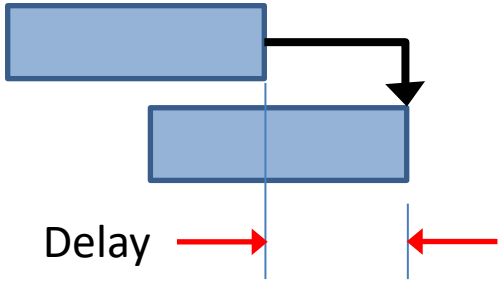


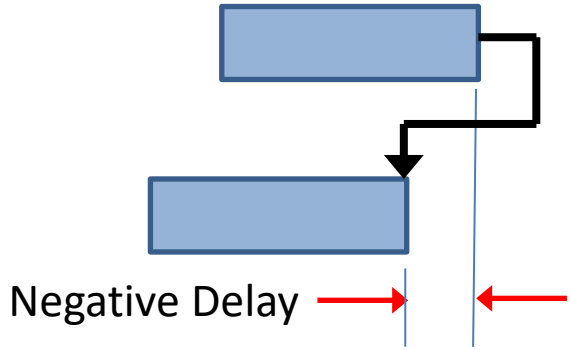
Dependencies

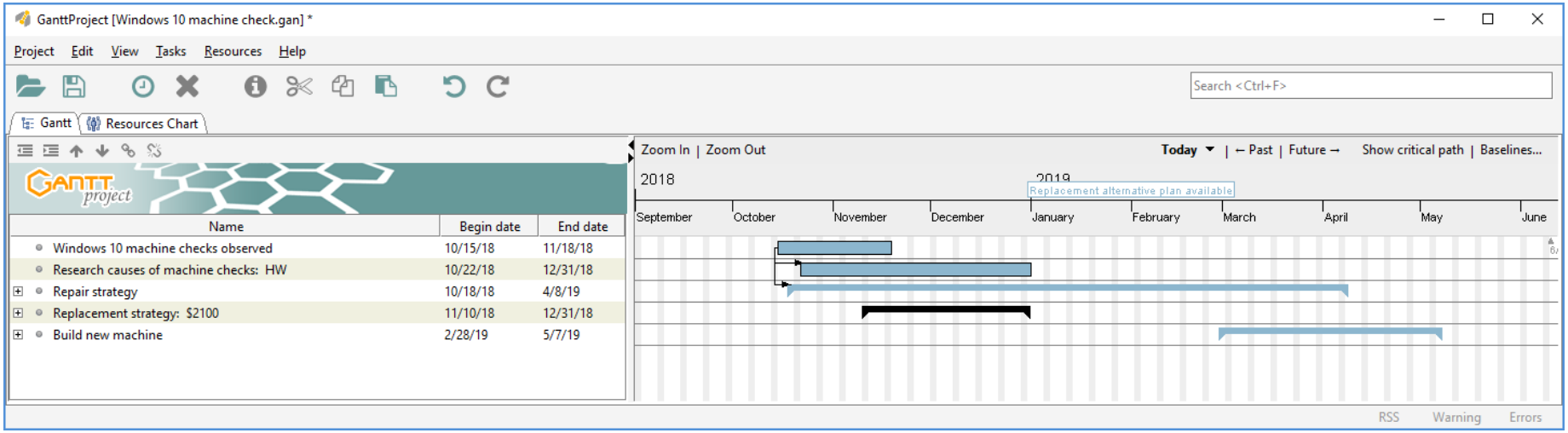
Finish-to-finish dependencies.



No delay






GanttProject [Windows 10 machine check.gan] *
Project Edit View Tasks Resources Help
Search <Ctrl+F>
Gantt Resources Chart
Zoom In | Zoom Out
Today | ← Past | Future → Show critical path | Baselines...
2018
2019
Replacement alternative plan available
September October November December January February March April May June
Name Begin date End date
Windows 10 machine checks observed 10/15/18 11/18/18
Research causes of machine checks: HW 10/22/18 12/31/18
Repair strategy 10/18/18 4/8/19
Replacement strategy: $2100 11/10/18 12/31/18
Build new machine 2/28/19 5/7/19
RSS Warning Errors

GanttProject [Windows 10 machine check.gan]

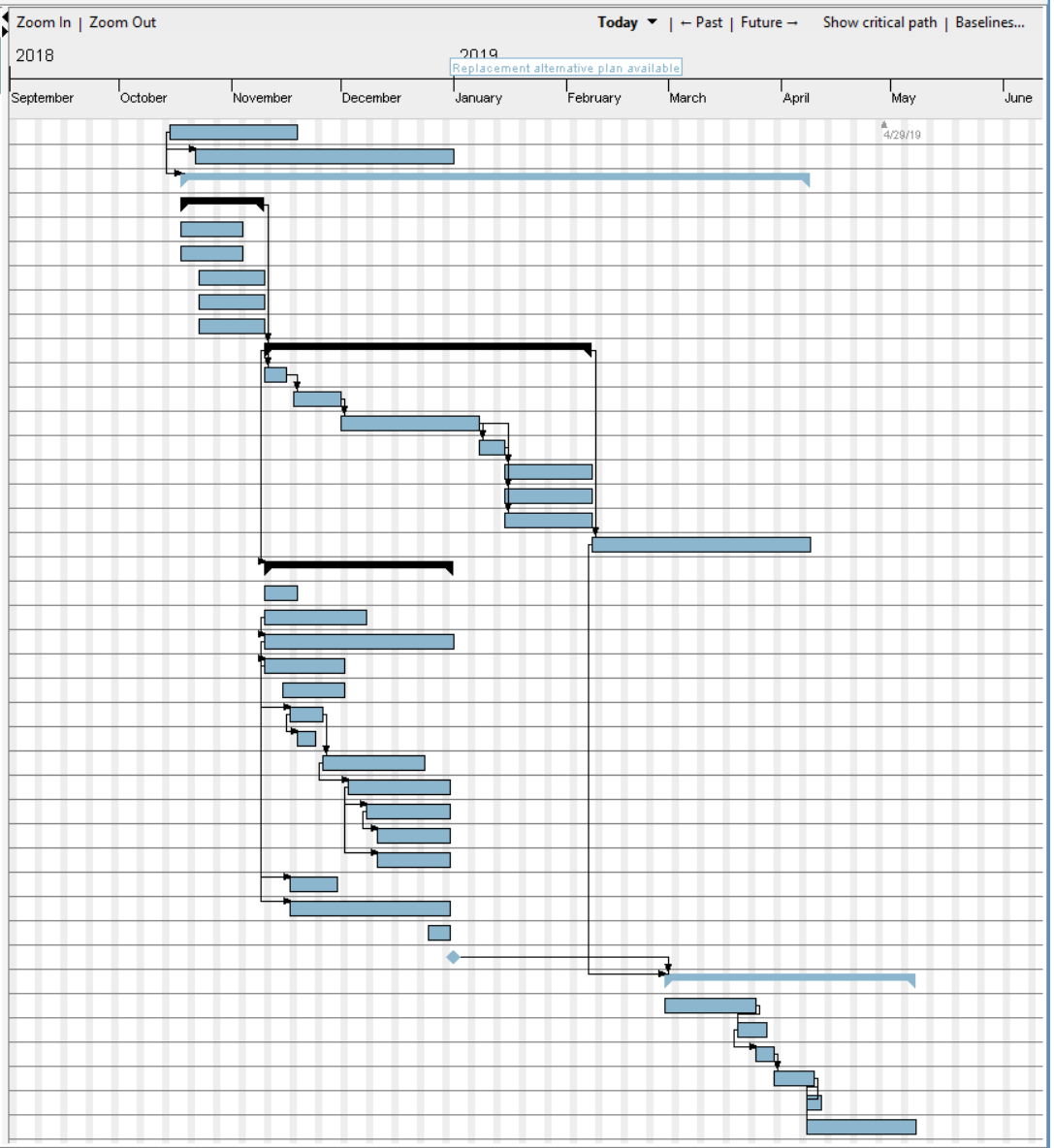
Project Edit View Tasks Resources Help

Search <Ctrl+F>

Gantt | Resources Chart

Zoom In | Zoom Out Today | ← Past | Future → Show critical path | Baselines...

Name	Begin date	End date
Windows 10 machine checks observed	10/15/18	11/18/18
Research causes of machine checks: HW	10/22/18	12/31/18
Repair strategy	10/18/18	4/8/19
Check for simple problems	10/18/18	11/9/18
Check for clogged or loose CPU fan	10/18/18	11/3/18
Check for loose connectors	10/18/18	11/3/18
Run Windows 10 integrity check (sfc /scannow)	10/23/18	11/9/18
Disable or remove non-critical I/O devices	10/23/18	11/9/18
Run memory tests	10/23/18	11/9/18
Begin swapping parts: $410	11/10/18	2/7/19
Decide to buy a new PS	11/10/18	11/15/18
Replace power supply $140	11/18/18	11/30/18
Decide to buy a new motherboard	12/1/18	1/7/19
Decide to buy a new CPU	1/8/19	1/14/19
Replace motherboard $180	1/15/19	2/7/19
Replace processor $45 cpu + $20 thermal grease	1/15/19	2/7/19
Replace case fan $25	1/15/19	2/7/19
Evaluate repaired machine	2/8/19	4/8/19
Replacement strategy: $2100	11/10/18	12/31/18
Risk assessment: System restore	11/10/18	11/18/18
Define need: High end Windows Intel desktop for ML	11/10/18	12/7/18
Define requirements: Intel, Asus MB, Lian Li case	11/10/18	12/31/18
Identify processor options: Intel Core i9-9740x vs i9-9900K	11/10/18	12/1/18
Identify likely better choices in 6 months: 10nm	11/15/18	12/1/18
Understand 8th vs 9th gen Intel Core series	11/17/18	11/25/18
Understand the Meltdown and Spectre vulnerabilties	11/19/18	11/23/18
Decide 8th vs 9th generation strategy: 9th Gen	11/26/18	12/23/18
Identify best CPU available now: Intel Core i9-9900K $530	12/3/18	12/30/18
Pick a motheboard: Asus ROG Mqximus XI Code Z390	12/8/18	12/30/18
Pick memory: 64GB Corsair Vengeance DDR4 3200	12/11/18	12/30/18
Pick cooling; Corsair Hydor H80i V2 water $95	12/11/18	12/30/18
Pick an M.2 storage card: Samsung 970 EVO $497	11/17/18	11/29/18
Pick a case; Lian Li PC-8NB $89	11/17/18	12/30/18
Pick a DVD Burner $23	12/25/18	12/30/18
Replacement alternative plan available	12/31/18	12/31/18
Build new machine	2/28/19	5/7/19
Order parts	2/28/19	3/24/19
Assembly	3/20/19	3/27/19
Configuration	3/25/19	3/29/19
Install Windows 10	3/30/19	4/9/19
Restore user files	4/8/19	4/11/19
Install applications	4/8/19	5/7/19



RSS Warning Errors

An exercise

You've been invited to a job interview in Mountain View, CA. They've given you several possible dates and asked that you make your own travel arrangements, which they will reimburse. Draw a sensible Gantt chart of the following activities.

Book a flight and a hotel.
Catch an Uber to the airport.
Decide what day is best.
Pack your bags.

To keep your chart simple and easy for us to grade, draw it simply as boxes and arrows with the one-word name of the activity ***inside*** each box.

Spoiler alert

The next page contains a solution.

Did you get something like this?