Queue Scheduler Scoring and Metric

2022 April 14

Introduction

Every scheduling algorithm requires that the objects being scheduled be given a value --- often called a priority, rank, or score --- that will be used to compare and pick the objects for the schedule. In multi-objective problems there may be more than one value. In many cases the goal of the scheduler is to maximize the cumulative value of the final schedule. This document describes the current work and future plans on how this value, hereafter called the ‘score’, will be calculated for the Gemini adaptive queue scheduler being developed as part of the GEMMA TDA project. The scoring algorithm is mostly independent of the scheduling algorithm and, since it attempts to codify policy, it is crucial to creating observing plans that meet all the observatory’s goals.

Score

The scoring algorithm is an attempt to codify the criteria that Gemini queue coordinators (QCs) have used over the years for creating their manual plans. The criteria include the observatory program completion goals and guidelines for handling different TAC ranking bands, different types of programs, targets of opportunity (ToO) observations, timing windows, conditions, etc. following the guidelines given to the QCs. The scoring algorithm that we are currently considering is

Score = pre-imaging^{a} × (internal priority)^{b} × (conditions match)^{c} × metric^{d} × visibility^{e} × (hour angle weight)^{f}

where a–f are exponents. The following sections describe the details of each term. The exponents are all 1 by default but are included so that the relative contributions of each term can be adjusted. The aim of this work is to determine the minimum number of terms and their exponents that will meet the QC criteria and allow the scheduler to create robust queue plans that are at least equivalent to manually-created plans.

Pre-Imaging

Multi-object spectroscopy (MOS) observations often require pre-imaging observations taken a few weeks in advance of the spectroscopy so that the masks can be designed, fabricated (cut), and shipped. Timing constraints may be sufficient to ensure that these are done sufficiently early but if necessary these will be given an extra boost in score.

Internal Priority

The legacy OCS allows PIs to set a Priority of Low, Medium, or High on each observation. This is supposed to show relative priorities within the program and is mostly used by QCs as a tiebreaker between otherwise similar observations in the program. This is a useful capability to maintain so several options have been discussed. The implementation is a bit complicated since any changes to the score are global, it affects the comparison with all other observations. The Las Cumbres Intra/Inter-Proposal Priority (IPP) system was reviewed but it appears complicated and confusing and it can lead to PIs trying to game the system. We will start by investigating the following two options:

1. The QPT currently scales the global observation score by priority after evaluating the priorities of the other observations in the program.
2. The scoring algorithm could include a second pass in order to identify observations in the same program at similar positions on the sky whose natural scores do not agree with the requested priorities. Then the scores are adjusted to match the priorities by swapping or scaling. This effectively accomplishes what the QCs do now with QPT, so this is the preferred approach.

Conditions Matching

A fundamental principle is that observations can only be done when the conditions are at least as good as the constraints. Therefore, the conditions constraints are compared to the actual conditions and the score is set to zero if the conditions are not good enough. The actual conditions can be a function of time, so the scheduler can support forecasts. If no forecast is available, then the current conditions are assumed for the rest of the night. The actual conditions also include wind speed and direction so that the scheduler can support any telescope pointing restrictions.

Observing in conditions that are much better than necessary should be discouraged in order to reserve that time for observations that really need the better conditions. At the moment this is done by reducing the score of observations in conditions that are better than needed by a factor related to the percentile difference. For example, in IQ70 an observation that requires IQ85 will have its score reduced by 1.0 - (0.85 - 0.7) = 0.85. Each condition is treated separately and the reductions are cumulative.

Observations that are setting early in the night need extra priority so that they can be observed before they become inaccessible. These observations should have a high visibility fraction. In addition, the ‘conditions better than needed’ penalties mentioned above are not applied if the hour angle of an observation is positive (indicates setting) at evening nautical twilight.

Metric

The metric is used to quantitatively evaluate Gemini’s queue scheduling and execution over a period of time, typically a semester. It encodes the high-level observatory aims (see the Scheduler Science Requirements document) such as completeness goals, relative band importance, etc. It is designed to replace the completion rates that we use currently. The high-level job of the scheduler will be to maximize the metric over a period of time.

At the time of this writing there is no single number, or metric, that is used to evaluate operations. In general we've used program completion (the fraction complete to some minimum level) to evaluate how we are doing. Analysis of completion and publication statistics shows that once you get much above 60%, the likelihood of publication doesn't change. In recent semesters we have used 80% completion as a performance metric in order to be comfortably above 60%. Some statistics are given on the Completion Expectations web page. While the priority for a program can drop once the completion rate is above 80%, the goal should still be to complete as many programs as possible.

Several forms of the metric have been investigated and are described below.

Step Function

A simple metric consistent with the top-level aims document is a step function that rewards programs for reaching the "publication level" (currently 80%) and Band 1 programs that are completed (Figure 1).

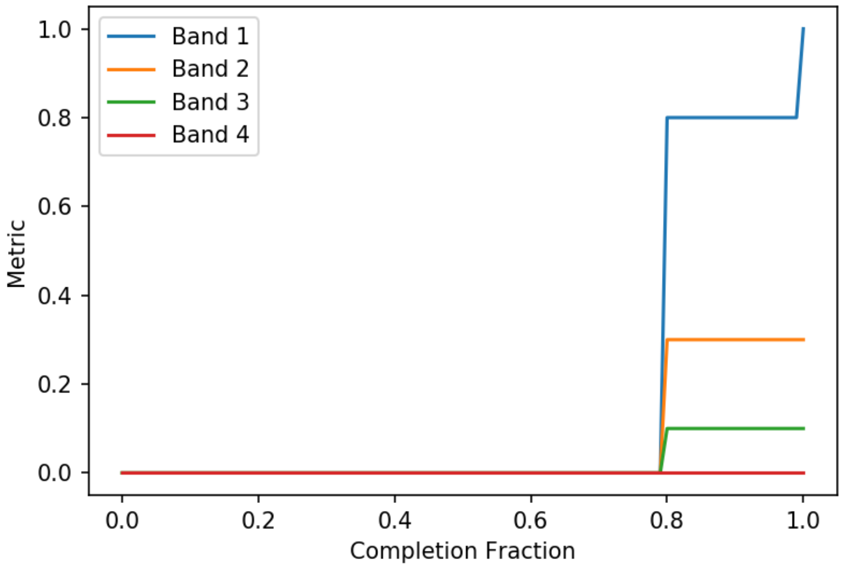


Figure 1: A metric that only counts programs that reach the "publication level" and Band 1 programs that are completed.

Linear Metric 2

A weakness of the step metric is that it does not provide an incentive or guide for how the scheduler should reach the completion goals. Also, data taken for partially completed programs are still valuable and the observatory should get some ‘credit’ via the metric for this work. This is especially true for programs that just miss a threshold, e.g. 75% completion in Band 2.

Therefore, we developed a linear metric with a change of slopes at the 80% completeness threshold (Figure 2). The metric has different slopes for the different bands and there can be a change of slope and a “bonus” at the completion threshold as a way of encouraging programs to reach this level. Band 1 also has a program completion bonus.

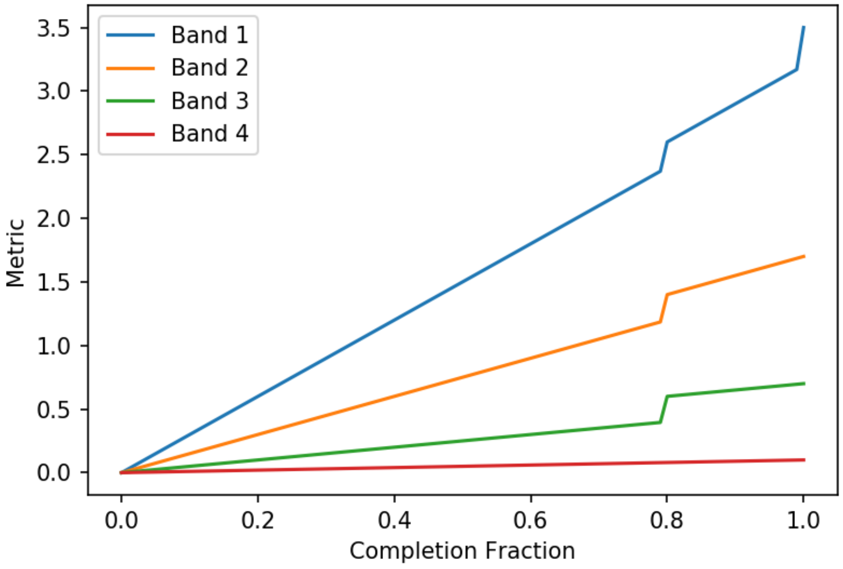


Figure 2. A linear metric with bonuses at the completion threshold and for Band 1 program completion.

Parabolic Metric 1

In the next iteration the importance of programs with lower completion is reduced and the influence of programs that are close to the completion goals is increased. This is done with a parabolic (2nd order polynomial) function below 80% completeness and a linear function above (Figure 3).

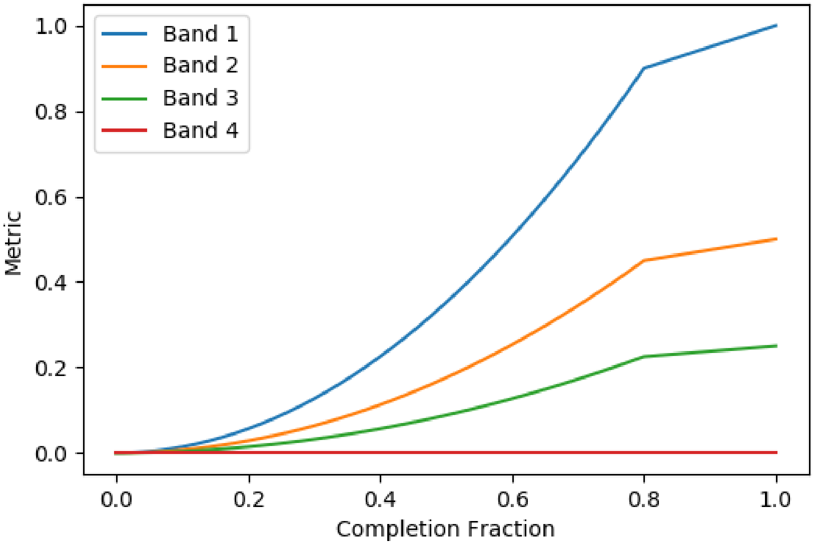


Figure 3. A parabolic/linear metric.

Parabolic Metric 2

With all the previous metrics a lower-band program with relatively high completion will have a higher metric than a high-band program with a lower completion. This could drive the scheduler to try to finish a lower-band program rather than start a higher-band program. We can avoid this by completely separating the bands. A parabolic/linear metric that does this is shown in Figure 4.

The observatory also aims to help thesis students by prioritizing their programs. Therefore, this metric also includes an additive bonus for thesis programs (dashed lines). The use of this bonus was discussed with the Gemini participants at the February 2022 Operations Working Group meeting. Noone objected and a few participants were enthusiastic about this since they already give thesis programs a ranking boost.

This is the currently preferred metric and the one with which most prototype scheduler tests have been done.

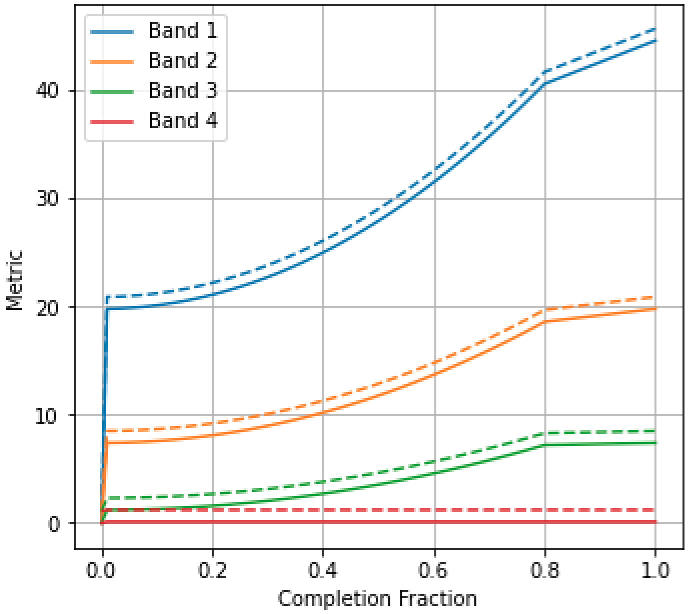


Figure 4. A parabolic/linear metric with complete band separation. Once started, any higher-ranked program will have a higher metric than a lower-ranked program. The dashed line is a small additive bonus for thesis programs.

The concept for the use of the previous metrics is to push programs “up the curve”, not to integrate the metric under the curve. If a scheduler more naturally maximizes the integral of the metric then the metric function, especially the band-separating parabolic function (Parabolic Metric 2), may need to be modified.

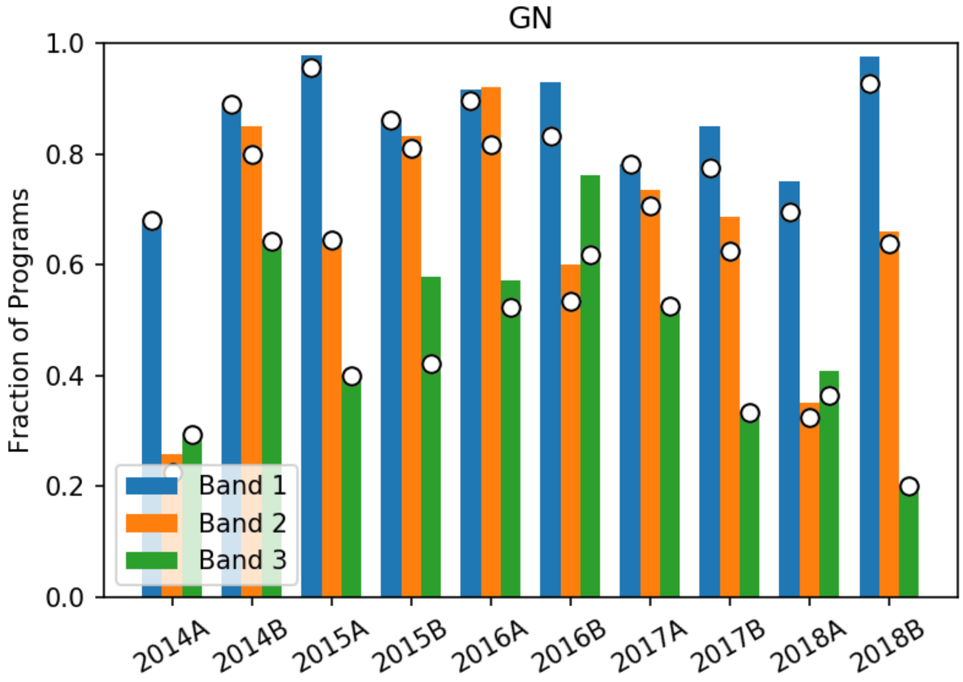
Comparison with Standard Completion Statistics

In order to evaluate how the metrics can be used to evaluate full semesters, we apply the metrics described above to some historical time accounting data in order to compare the results with the standard completeness statistics.

In order to make the results comparable to those on the public web pages we choose GMOS(N/S), F2, and GNIRS non-ToO observations from queue programs as representative of optical and infrared observations using facility instruments. Completion fraction is program time divided by the allocated program time, but it is set to 1 if the program is marked as complete. We run the test using 5 years of data from 2014-2018.

The standard completion statistics are given in the top panels of Figures 5 and 6. The length of each bar gives the fraction of programs in a semester/band that reaches at least 80% completion. The white circles show the fraction of programs that are 100% complete. The results from the metrics are shown in the bottom panels of Figures 5 and 6. The metrics are all normalized to the same maximum value to make them easier to compare. The overall trends are similar. As long as the metric separates the bands, the exact form of the metric does not have a significant effect, the completion of Band 1 programs dominates.

Therefore, metrics such as those presented can be a useful measure of the overall productivity of the observatory in a given semester. The differences between semesters are mostly a result of weather loss, shutdowns, and other extraneous factors. The metric per science hour could be used as a quantitative measurement of the efficiency of queue operations.



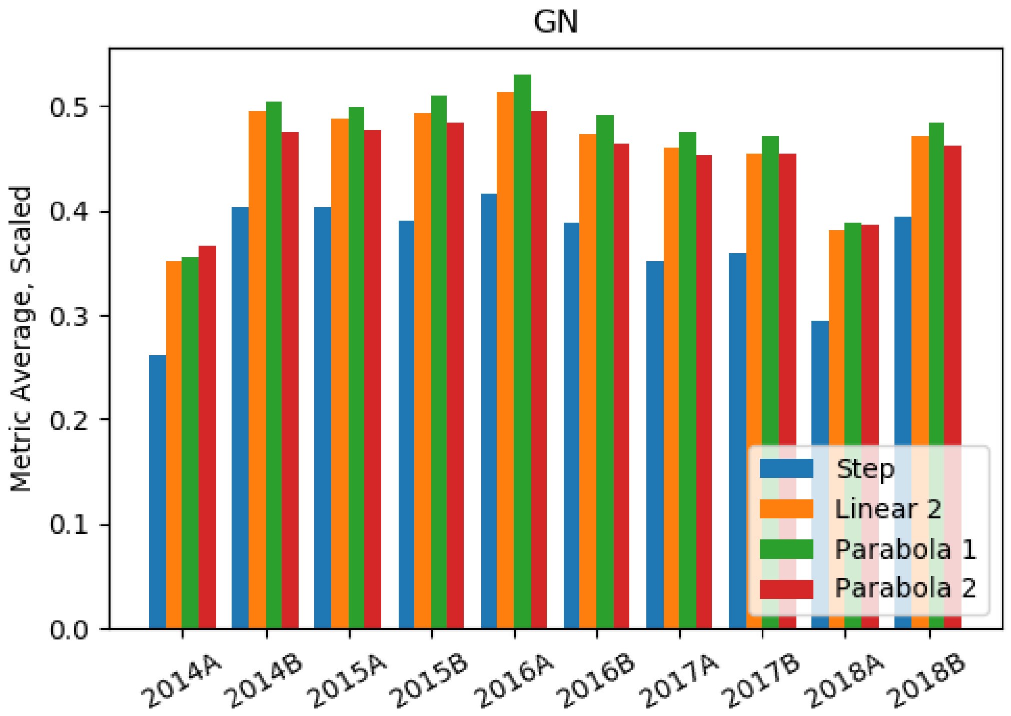


Figure 5. Top: Standard completion statistics separated by Band. The length of each bar gives the fraction of programs in a semester/band that reaches at least 80% completion. The white circles show the fraction of programs that are 100% complete. Bottom: The results from the four metric options for Gemini North.

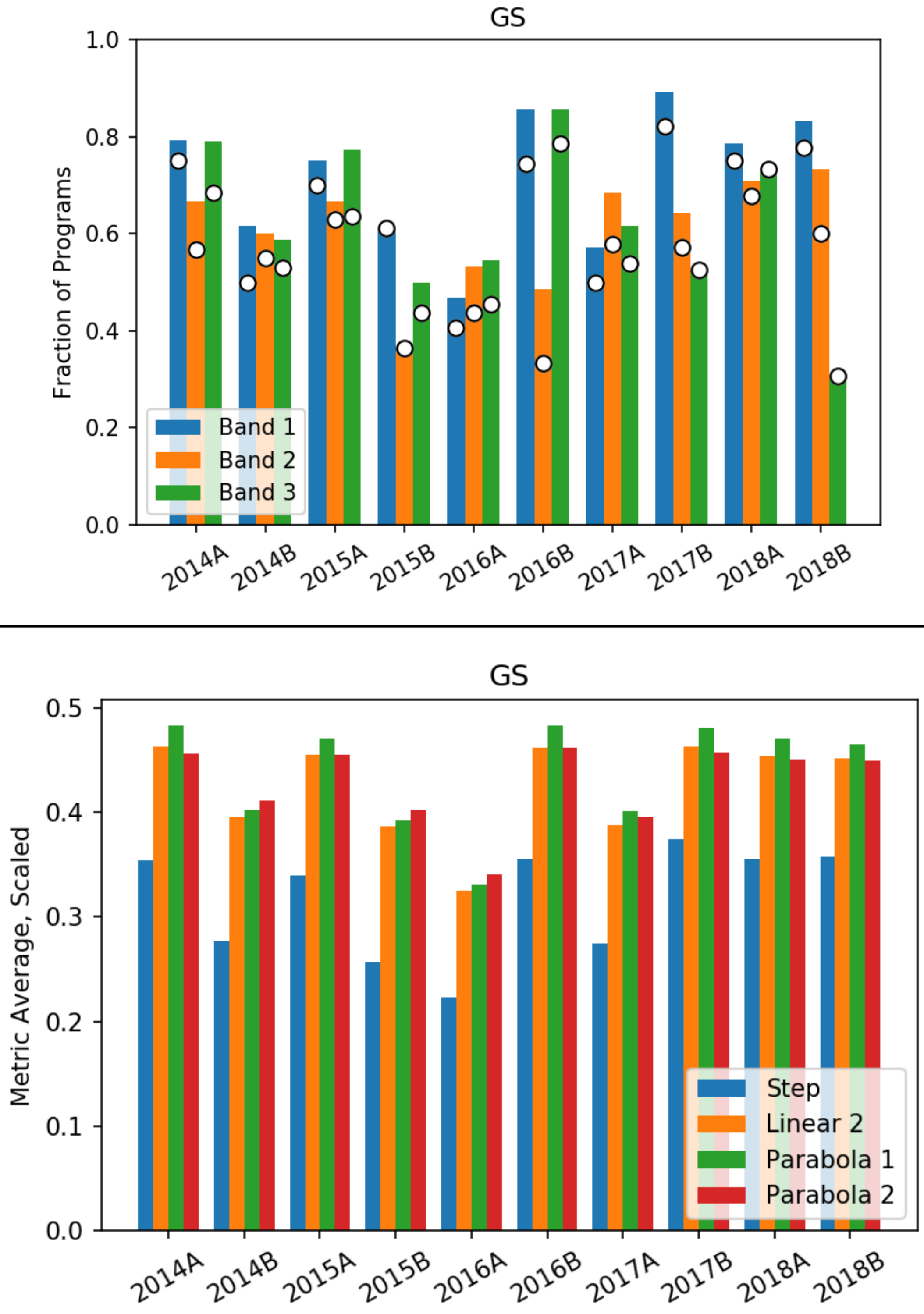


Figure 6. As Figure 5 but for Gemini South.

Visibility Fraction

The visibility fraction (visfrac) is the length of an observation or group divided by the total time that it is available to be scheduled. The denominator will include airmass/hour angle and sky brightness constraints, user-defined timing windows, and telescope/instrument/component availability. If the telescope schedule is not defined yet then we will use historical statistics to estimate instrument and component availability. For example, we can determine the fraction of the time that the GMOS-N R831 grating was installed over the last three semesters. The visibility calculation is made from the date of interest until the end of the program's active period. Fast-Turnaround (FT, submissions once a month and reviewed by the proposers) programs are active for three months and other programs are active for their natural semester while appropriately accounting for "fuzzy" semester boundaries and Band 1 persistence (see Appendix A). Therefore, targets that are setting or have stringent timing constraints will have a larger visibility fraction since the total available time (denominator) is smaller. Likewise, FT programs get a slight score boost since they are active for a shorter time. This also gives long observations/groups a boost since the numerator is larger.

This scheme works as expected for the single-site case but has a bias in the multi-site case since the score will be systematically higher at the site where the observation is *less* visible (has a lower maximum elevation). While all constraints are still met with the current algorithm, the scheduler should try to schedule observations at the site where the target is higher. We will start by considering the following options that can improve this behavior, but these need prototyping and use in simulations to determine the effectiveness:

- Include the total available time for all the sites in an OR group with the same target. This results in a smaller visibility fraction and lower score at all sites. Using the same denominator for all sites will make the scores more uniform and reduce the bias towards the less-optimal site. However, traversing group trees and managing the bookkeeping may be complicated.
- Include an airmass or elevation factor such as dividing the visibility fraction by the minimum airmass. This has the advantage of being independent of the group tree and it worked as desired in some initial tests, but the effect on all observations with high minimum airmass needs to be evaluated.

Hour Angle

The hour angle is used to encourage the scheduler to place observations when they are near the meridian, or minimum airmass. For targets with minimum zenith distances of less than 40 degrees the maximum weight occurs at HA = +1, or just after transit (solid curve in Figure 7). This is an attempt to avoid tracking complications at the meridian/zenith and give setting observations a bit more weight. However, for objects with minimum zenith distances greater than 40 degrees (airmass > 1.3) the weighting function peaks at HA=0 (at transit).

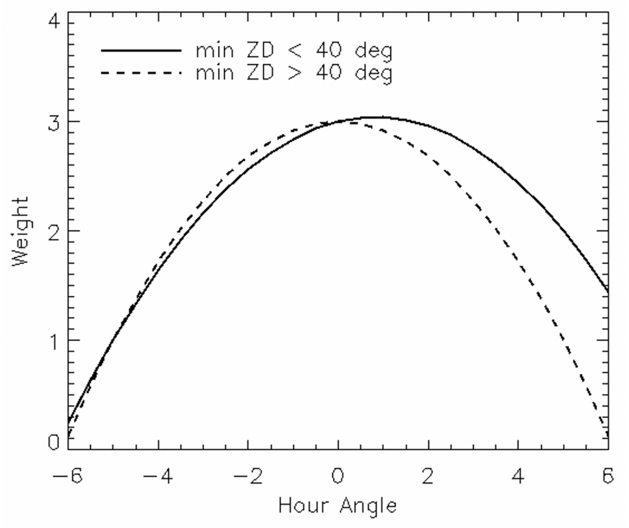


Figure 7. Hour angle weighting functions.

Finalizing the Scoring Algorithm

The scoring algorithm will be continually refined using the processes described in the Scheduler Testing and Verification document. We will use the validation mode of the scheduler to evaluate changes in the algorithm. Initial comparisons between QC plans and GreedyMax automatic plans show that the scheduler produces results that are quantitatively comparable to human-generated plans (see the Appendix of the Testing and Verification document). These tests have also revealed weaknesses in the scoring algorithm that have motivated the addition of new terms and changes in approach. While some improvements are trial and error, we are also planning a statistical analysis of historical queue plans in order to determine how the QCs generate their plans and the measure coefficients (exponents) on the terms. This may make use of principal component analysis and machine learning classifiers.

Once in operation, the active scheduling algorithm will be described on Gemini’s public web pages for transparency.

Appendix A

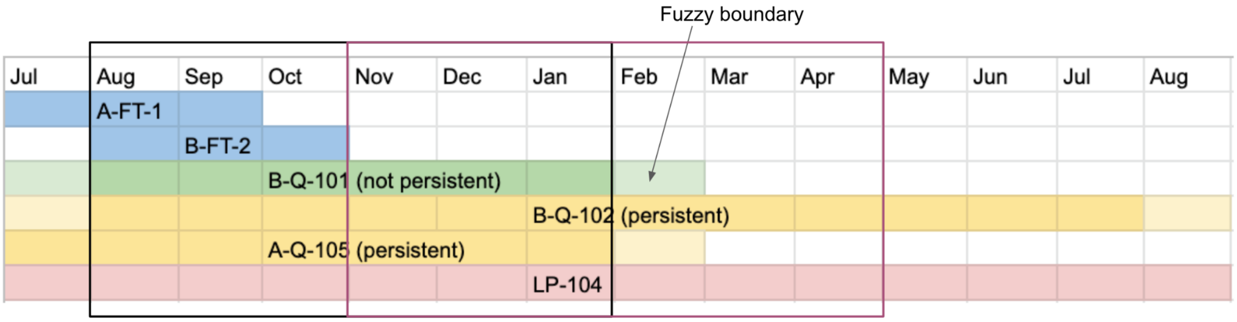
Here we describe an option for handling visibility fraction calculations for long-duration programs (e.g. persistent Band 1 and LLP) in order to meet the scheduling goals. These long-duration programs should not be put off, or there is the risk that they won’t get completed. Also, if the community perceives that these programs get lower priority, then they are disincentivized from applying.

In this example, LLP programs are given a single program id for the life of the program, unlike in the legacy OCS where there is a new program id each semester. In this case we must try to complete each semester's observations/allocation, we should not put them off because the program will continue.

In addition, in some cases it may be appropriate to prioritize non-persistent programs depending on the relative completeness fractions, etc.

In this approach the visibility fractions (observation length / sum of visible hours) are calculated from the current date to some period (e.g. 6 months) into the future or until a program's natural end date (allowing fuzzy boundaries), whichever is earlier.

At the start of the semester all Band 1 programs are equivalent (see the black rectangle in the figure below).



As time progresses (red box in figure for a plan date of November 1) the visible hours for non-persistent Band 1 decrease, increasing the visibility fraction and therefore the score relative to persistent Band1 from the same semester, all other things being equal.

As each program reaches its end date, the visibility fraction, and therefore the score, rises naturally to push any remaining observations to completion.